Stats 140XP Final Paper: A Good Company

Carly Jeung, Donggyu Kim, Cherry Li, Shiyu Murashima,
Andrew Schweitzer, Amber Tsao, Michelle Wang

Winter 2024

Contents

1 Abstract

The job search process is an essential part of our lives, as we try to seek out the company that is best suited to our interests. Over the course of time, many companies have established great renown, while others continue to struggle to make a name for themselves. Thus, the question arises: What makes a company a good company? This study aims to address this question by examining the factors that allow a company's rating to stand out among others. We conducted an analysis on information found in Indeed job postings through various statistical modeling techniques, including random forest, CatBoost, multiple linear regression, and XGBoost. As expected, our results have shown that the most significant factor in determining a company's rating is salary. Other important factors include but are not limited to: license requirements, occupational category, and company location.

2 Introduction

2.1 Background

The data used in this study was provided by Indeed—an online platform that allows users to navigate millions of job listings worldwide. This particular dataset was originally used in the 2018 DataFest competition at UCLA, sponsored by the American Statistical Association. Each observation included a series of details about a single job listing posted to the job search site.

2.2 Description of Data

The table below shows the name, type, and description of the variables in the dataset.

Name	**Type**	**Description**
date	Character	The recorded viewing date of the job posting
companyID	Numerical	The ID of the company
jobID	Numerical	The ID of the job
country	Character	Country of job posting
stateProvince	Character	Name of the state/province of the job posting
city	Character	Name of city of job posting
avgOverallRating	Numerical	Average rating of the company (1-5 stars)
numReviews	Numerical	Total number of reviews the company had
industry	Character	Industry associated with the company
normTitle	Character	The normalized/canonical job title
normTitleCategory	Character	The area of association of job posting
descriptionCharacterLength	Numerical	Number of characters in job description
descriptionWordCount	Numerical	Number of words in job description
experienceRequired	Numerical	Minimum experience required for the job (in years)
estimatedSalary	Numerical	Estimated annual salary for the job
salaryCurrency	Character	The currency of the salary
jobLanguage	Character	The language the job required
supervisingJob	Factor	Whether this job is classified as a supervising job
licenseRequiredJob	Factor	Whether this job is classified as requiring a license
educationRequirement	Character	The education requirement for the job
jobAgeDays	Numerical	Age of job (in days), based on job creation date in Central Time; this resets if the job is 'refreshed'
clicks	Numerical	The total number of clicks on the job on the date
localClicks	Numerical	The total number of clicks on the job from a local user (same city and country) on the date

Table 1: Description of Dataset

3 Exploratory Data Analysis

3.1 Data Cleaning

Before conducting any analysis, our group first cleaned the Datafest 2018 dataset by dealing with the missing values, insignificant columns, inconsistent units, and repetitive observations. The original full dataset had 14,586,035 observations of 23 variables, and we cut it down to 199,417 observations of 17 variables.

First, we removed variables which were unimportant to our question of interest, including the "date", "descriptionCharacterLength", and "language" columns. We also removed the "normTitle" and "industry" columns since the majority of the values in the "industry" column were empty, and the "normTitleCategory" column provided more useful information regarding the occupational category.

To handle the missing values in the "numOfRatings" column, we removed all observations with an NA value, which corresponded to the observations with a value of 0 in the "avgOverallRating" column. As noted in our data dictionary, companies were rated on a scale of 1 to 5 stars, so a company with an average rating of 0 meant it had no ratings. Non-rated companies were not useful for our analysis since we used the average rating as a measure of how good the company was deemed to be. We also removed all observations with a blank value for the "city" and "normTitleCategory" columns since they were important variables of interest

To standardize the various units of currency, we converted all currencies of the estimated salaries into US Dollars using the average 2016 exchange rate of 1 USD = 1.25 CAD and 1 USD = 0.85 EUR. Before that, we also used the country origin of the jobs to fill in the currency for the columns with a missing currency type.

Finally, we noticed many repetitive observations sharing the same "jobId". Among the 5,400,000 observations that were left after handling missing values, there were only 199,417 unique job IDs. This repetition was due to most job postings having a separate observation for each day that had elapsed since its initial posting. We grouped the data by "jobId" and summarized using the maximum of "jobAgeDays" (i.e. total number of days the job has been posted) and the sum of "clicks" and "localClicks" (i.e. total number of clicks for each job posting).

3.2 Analysis of Key Variables of Interest

For our exploratory data analysis, we conducted an analysis on the following variables of interest: the state in which the company is located (stateProvince), the occupational category (normTitleCategory), the company's number of reviews (numReviews), and the estimated salary for the job in USD (estimatedSalaryUSD).

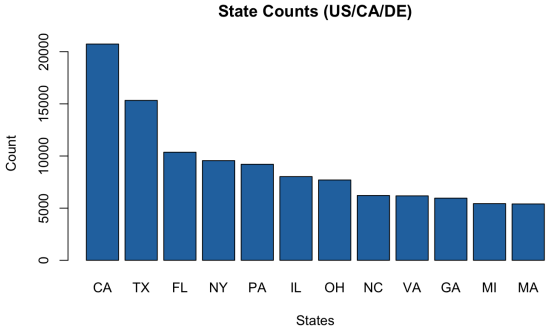

Figure 1: Distribution of Job Listings by State

In **Figure 1**, we see that California dominated in the amount of jobs listed, reaching a total of 20,000 job listings. Texas came second with 15,000 listings, and Florida was third with 10,000 listings. **Figure 2** shows that most companies had between 0 and 100 reviews, with a select few having more than 10,000 reviews (not depicted in **Figure 2**, which is a more focused shot of the distribution).

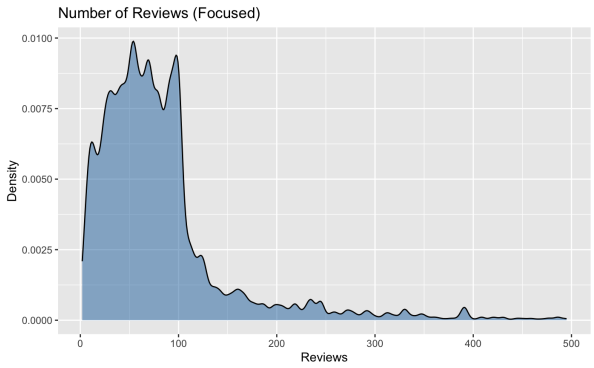

Figure 2: Distribution of Number of Reviews (Focused)

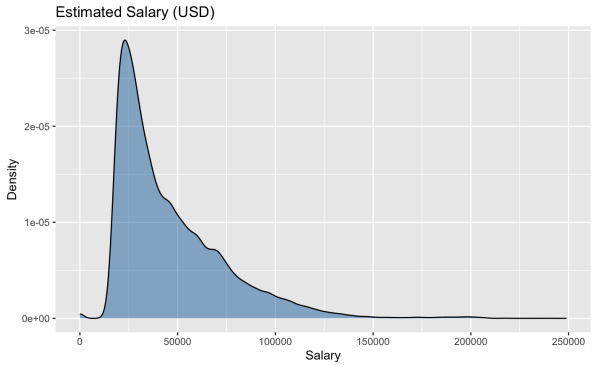

Figure 3: Distribution of Estimated Salary

Figure 4: Word Cloud of Occupational Category

In **Figure 3**, we see the estimated salary (in USD) across all job listings was around $25,000 to $30,000 on average. As expected, this distribution was skewed to the right, reflecting how most jobs were listed at the minimum annual salary of the time.

After conducting basic text analysis on the occupational categories (normTitleCategory) across all jobs, we produced a word cloud, as seen in **Figure 4**. The words "mednurse", "management", "retail", and "food" are highlighted (i.e. had the largest font size), demonstrating that these were the most common job titles.

3.3 Normalizing Rating by Amount of Reviews

Most importantly, we took a closer look at our response variable: the average rating of the company (avgOverallRating). In the cleaned dataset, the average overall rating was around 3.75 for all companies, as seen in **Figure 5**. We then used Bayesian estimation to normalize the rating by the corresponding number of reviews for each company. After this step, the average rating increased to around 4.0, with most companies in the 3 to 4 range (see **Figure 6**).

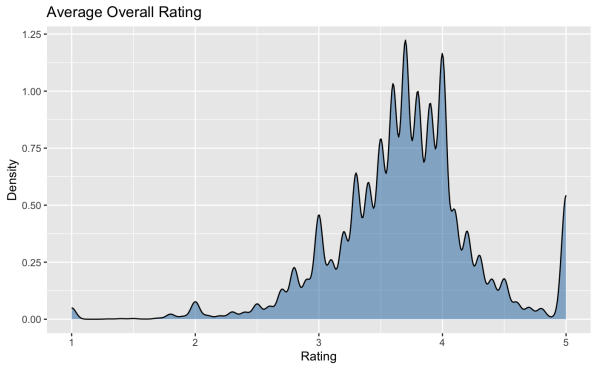

Figure 5: Distribution of Average Company Rating

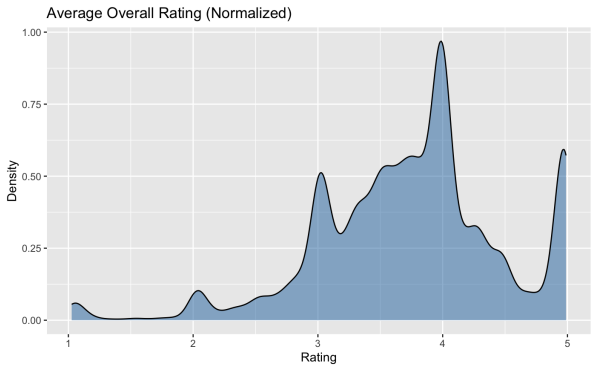

Figure 6: Distribution of Average Company Rating (Normalized

Figure 7 shows a heat map of the average overall rating by state in the U.S. Here, we see that Montana, Hawaii, and California were among the top states with the highest company ratings. On the other end of the spectrum, Pennsylvania had the lowest average overall rating of 3.55.

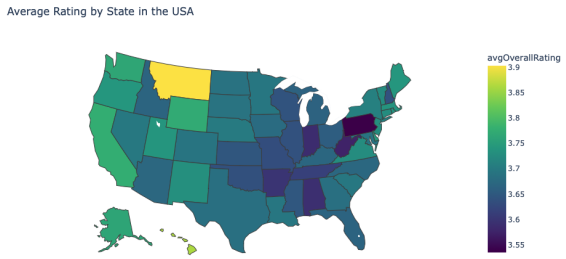

Figure 7: Heat Map of Average Company Rating by State in the U.S.

4 Data Modeling

To model our data, we applied the methods of random forest, CatBoost, multivariate linear regression, and XGBoost using the normalized company rating as the response variable. These models were chosen for their compatibility with categorical variables as well as their usefulness in identifying the most important variables in predicting company rating.

4.1 Random Forest

Excluding the company ID and job ID, a random forest model was run on the remaining 15 predictors using 500 trees and trying 4 variables at each split. In **Figure 8**, displayed below, the variable importance plot shows that the top 5 most important predictors were "estimatedSalaryUSD," "numReviews," "descriptionWordCount," "avgClicks," and "city." The resulting mean squared error (MSE) for this model was 0.292.

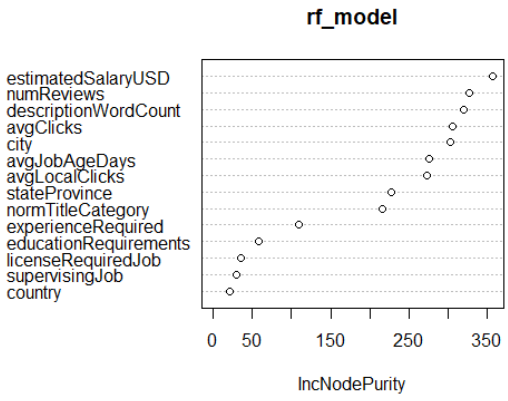

Figure 8: Variable Importance Plot for the Full Random Forest Model

Using these top 5 most important predictors, we ran another random forest model with only these variables as predictors. This simplified model used 500 trees with 1 variable tried at each split, and the resulting MSE was 0.304, which was slightly higher than the MSE of the full model. This indicates that the additional variables in the full model enhanced the prediction power of the model. However, although the full model was more accurate, the difference in MSE between the two models was small, so the simplified model still captured most of the variance despite using fewer predictors. Therefore, the simplified model was a good choice due to its nearly equivalent accuracy and the benefits of interpretation and computational efficiency.

4.2 CatBoost

Next, we tried a CatBoost model, where we first converted the categorical variables to factors. Using RMSE as the loss function, the top 5 most important variables were found to be "normTitleCategory," "stateProvince," "estimatedSalaryUSD," "educationRequirements," and "city." A feature importance graph displaying the importance of the other variables is shown in **Figure 9**.

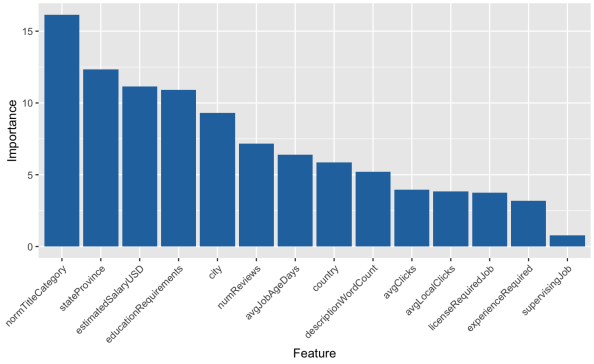

Figure 9: Feature Importance Graph for CatBoost Model

4.3 Multiple Linear Regression

We also tried multiple linear regression using "country," "stateProvince," "normTitleCategory," "descriptionWordCount," "experienceRequired," "estimatedSalaryUSD," "supervisingJob," "licenseRequiredJob," "educationRequirements," and "avgClicks" as predictors. Categorical variables were first turned into dummy variables during one-hot encoding. After running the model, we found that most of the variables were good predictors for the company rating. In particular, "estimatedSalaryUSD," "supervisingJob," "licenseRequiredJob," and "avgClicks" were very significant, as well as the "normTitleCategory" variable, although this was slightly more difficult to see since the features were split by each unique job title. We also checked the ANOVA and found the p-value to be $< 2.2\text{e-}16$, representing the significance of the overall model. The diagnostic plots for the model are shown below in **Figure 10**.

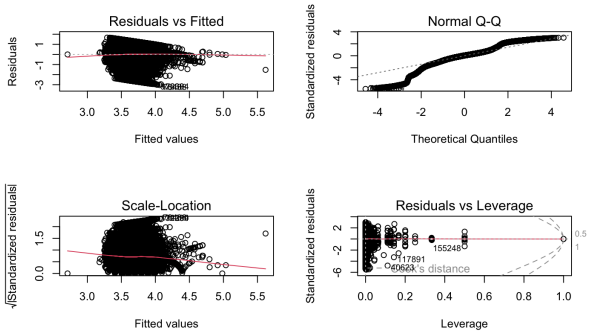

Figure 10: Diagnostic Plots for Multiple Linear Regression Model

We notice that the the Residuals vs. Fitted plot implies the variance is likely constant, and the Residuals vs. Leverage plot does not suggest outliers. However, the Normal Q-Q plot shows signs that the data may not be normal. In addition, the Scale-Location plot does display a somewhat random scatter, but due to the lack of a horizontal line, homoscedasticity cannot be assumed.

Next, an inverse response plot was run to improve this model. Using the optimal lambda of 2.272, we applied this transformation to the response variable and fitted a new model. **Figure 11** represents the results of the inverse response plot, and **Figure 12** shows the resulting diagnostics of the improved regression model.

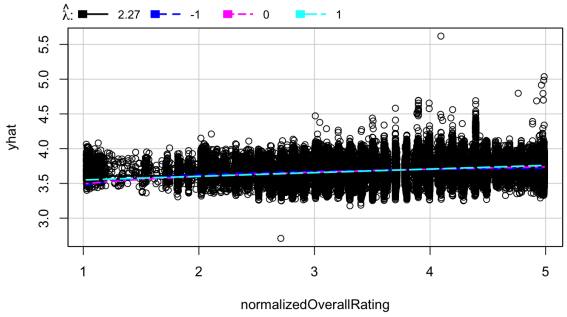

Figure 11: Inverse Response Plot

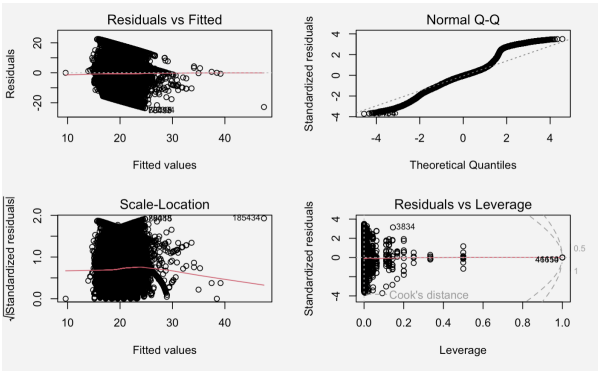

Figure 12: Diagnostic Plots for Improved Multiple Linear Regression Model

Once again, the condition of constant variance is satisfied and there are no outliers, and the Normal Q-Q plot also looks more normal. The Scale-Location plot looks better than before, but we still may not be able to assume homoscedasticity. We also attempted to use variable selection using the adjusted R-square, BIC, and forward and backward AIC methods, but the results were not ideal.

4.4 XGBoost

Lastly, we ran an XGBoost model as an alternative method to determine feature importance. Although it was difficult to measure the importance of most of the predictors due to the fact that one-hot encoding split the features of the categorical variables into separate values, the predictor "estimatedSalaryUSD" was found to be the most important by far, with "licenseRequired" coming in second. Both of these predictors were also found to be significant in the MLR model, and the high importance of the "estimatedSalaryUSD" variable also aligns with the result from the random forest model.

5 Conclusion

5.1 Results

This analysis sought to identify the key factors that influence a company's rating through the utilization of statistical models such as random forest, CatBoost, multivariate linear regression, and XGBoost. To summarize our findings, the most important variables from each model were:

- Random forest: estimatedSalaryUSD, numReviews, descriptionWordCount
- CatBoost: normTitleCategory, stateProvince, estimatedSalaryUSD
- Multiple linear regression: estimatedSalaryUSD, supervisingJob, licenseRequiredJob
- XGBoost: estimatedSalaryUSD licenseRequired

With the exception of the CatBoost model, all models displayed "estimatedSalaryUSD" as their most important feature, although this variable was still in the top 3 for CatBoost. In addition, "licenseRequiredJob" appeared in the top 3 for two of the models.

To further inspect characteristics of a good company, we obtained data from a poll by Gallup-Healthways showing which states had the highest and lowest well-being scores in 2017. Well-being scores, which include information on a person's sense of purpose, social relationships, financial condition, community involvement, and physical health, can affect a person's perception of a company, so we wanted to check for any correlations between these scores and company rating. We found the correlation coefficient to be approximately 0.199, suggesting a weak but positive correlation between company rating and well-being score.

5.2 Conclusion and Limitations

Despite these promising findings, this study has limitations that must be acknowledged. For example, although salary was found to be the most significant factor in a company's rating, a more detailed analysis is needed to verify its significance. Comparing salaries within the same job roles across different companies could lead to a better understanding of this variable. Another issue that arose was the presence of many companies without any ratings. Unfortunately, this study was unable to provide much insight on these small companies and startups that were unrated. Additionally, our analysis relies on data from specific countries, which may not fully represent the global context. It is also crucial to note that a high number of positive reviews online does not necessarily equate to a "good" company. Instead, future research could benefit from text analysis conducted on employees' reviews in order to more accurately determine their sentiments toward the company's quality. Finally, this dataset only involved data taken from Indeed, but ratings from other sites can also be brought in for more comprehensive information about a company's overall rating.

Despite these limitations, the results of this study are still useful for identifying characteristics of a good company. Overall, when considering the factors that make a company "good", the most important feature to look at is the salary of its posted jobs. Other factors that play a role include whether a license is required for a job, the occupational category of the job, and the geographical location of the company. Given the importance of finding a good company to work for, keeping these factors in mind will lead to higher satisfaction in the workplace.

6 Acknowledgements

Dataset from Indeed.com, 2018 ASA DataFest at UCLA
Gallup National Health and Well-Being Index 2017