Factors Impacting Number of LinkedIn Job Applications per Post View

JRLK - Jessie Ringness, Rebekah Kim, Laura Cai, Karen Dong

2023-11-14

Introduction and data

LinkedIn is a popular platform that connects companies and professionals spanning various levels of experience, and there are thousands of active job postings available on LinkedIn. Moreover, online job search services and platforms are now considered equally important for people to access a wide variety of opportunities compared to in-person job postings. With the sheer amount of postings, applicants may be overwhelmed by the vast amount of postings, and are less likely to come across some postings over others. Our primary research question is - what variables about job postings increase popularity among applicants?

This data set was created by Arsh Koneru-Ansari in August 2023, who used Python to scrape data directly from linkedin.com. The data contains jobs posted between August 10 and August 24, 2023. The scraper code is published in their GitHub.

The data dictionary for the variable definitions can be found in the ReadMe for the data. The variables we will focus on are:

- **`applies`:** number of applications that have been submitted
- **`views`:** number of times the job posting has been viewed
- **`max_salary`:** maximum salary offered in the position
- **`remote_allowed`:** whether job permits remote work (1 = yes)
- **`follower_count`:** number of company followers on LinkedIn
- **`listed_time`:** time when the job was listed, in UNIX time
- **`formatted_experience_level`:** job experience level (entry level, associate, mid-senior level, director, executive, internship)

- `type`: type of benefit provided (Medical insurance, Dental insurance, 401(k), Paid maternity leave, Disability insurance, Vision insurance, Tuition assistance, Pension plan, Child care support, Commuter benefits, Student loan assistance)

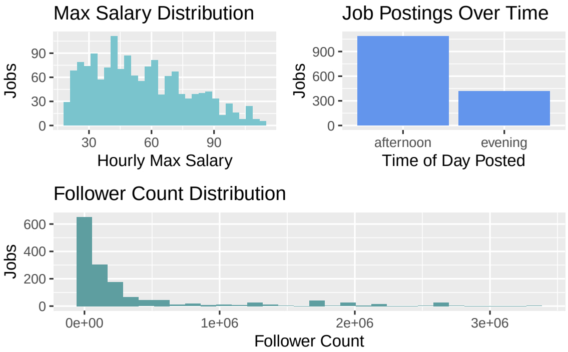


Fig 1.1. The distribution of hourly maximum salary is right-skewed with jobs in the data set having a generally lower hourly maximum salary. Given the apparent skewness the center is the median hourly maximum salary of \$50.80 per hour. The IQR describing the spread of the 50% of the distribution is \$33.94 per hour, demonstrating that the variability of the hourly maximum is relatively high. The histogram shows the middle 90% of the data to filter out the significant outliers.

Fig 1.2. The majority of posts were posted in the afternoon, with the next most popular time to post as the evening. 1085 jobs were posted in the afternoon, and 415 were posted in the evening. In the original data set, only a few jobs were posted in the morning and night: 4 jobs were posted in the morning and 7 were posted at night.

Fig 1.3. The distribution of follower count is right-skewed with jobs in the data set having a generally lower number of followers. Given the apparent skewness the center is the median of followers of around 70 thousand followers. The IQR describing the spread of the 50% of the distribution is over 280,000 followers, which means the variability of follower count in the data set is relatively high.

We saw the potential for an interaction effect between the maximum hourly salary and the required experience level of the candidate for the position, as typically those with less experience are paid less. We also saw the potential for interaction effects between hourly maximum salary and if remote work was allowed; however, the best-fit lines for there being no remote work and remote work parallel, indicating no potential relationship.

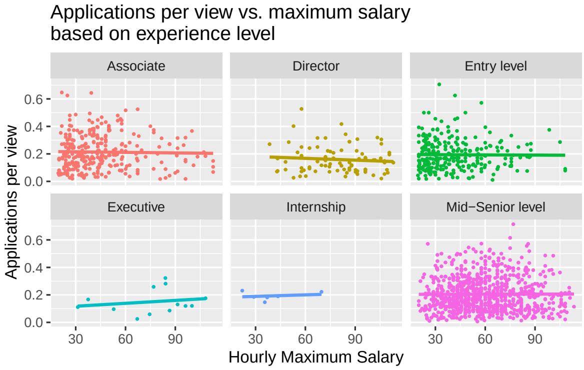


Fig 2. As the maximum hourly salary increases, the percentage of applicants out of viewers increases at different rates between different levels of experience required, although there is no apparent direction or shape between the hourly maximum salary and applications per view for each of the experience levels. The correlation between these two variables is around -0.0298, indicating the relationship is moderately weak. The distribution of applications per view based on experience level and maximum salary is mostly concentrated when the applications per view is less than 50% and the hourly maximum salary is about $60 for each experience level. The mid-senior level has apparent outliers when the hourly maximum salary is greater than 200. There is also not as much data for some of the experience levels, including internship and executive.

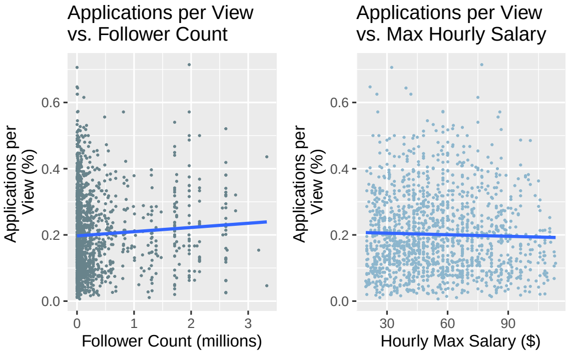


Fig 3.1. There is a possible linear correlation between follower count of a company and percentage of viewers who apply to the job. Most of the observations are concentrated to have less than 1,000,000 followers, although there are a significant number of observations above 1,000,000. The observations were filtered so that only the middle 90% of follower counts were observed, avoiding major outliers with too few or too many observations.

Fig 3.2. There is a slight positive linear correlation between a job's adjusted hourly maximum salary and percentage of viewers who apply to the job. Most of the observations are concentrated to have less than $200,000 for adjusted hourly max salary, so it would be beneficial to

filter out outliers that are above this threshold. Additionally, it would be beneficial to remove the outliers where 100% of viewers applied to the job (`per_applies` = 1.0)

Methodology

Intro

The data provides information about job listing details such as views, applications, time posted, etc., as well as whether or not certain benefits were offered in the job listing on jobs from August 10-24, 2023. This information was scraped from LinkedIn between August 23 and 24, 2023.

Joining Datasets

While most of our data is from the `job_postings` data set, we also wanted to include employee and follower count from the `employee` data set, and the type and number of benefits listed from the `benefits` data set. We joined all data sets together by `company_id`, and saved the data set as `linkedin`.

Benefits

Then, we added the number of benefits to create a new `tot_benefits` predictor, which tells us the total number of benefits listed. If `remote_allowed` and `tot_benefits` had NA entries, we assumed the job did not allow remote work and did not list any benefits. Consequently, we imputed them to 0's. However, the majority of entries in `benefits` are `NA`, which means no benefits were scraped from the listing. To make data from `benefits` a useful predictor, we created a new categorical variable `if_benefits`, which tells us if any or no benefits are listed. If any benefits (i.e. paid maternity leave or 401k plan) were listed, then the post is considered `listed`, and otherwise considered `none`.

Salary

We also made assumptions about other variables to normalize predictor variables. To compare salaries even if they were listed in different formats (such as hourly pay, monthly, or yearly salary), we normalized the variable using the categorical variable `pay_period`, which tells us if the job pays its worker the `max_salary` or `min_salary` amount, with hourly, monthly, and annual payments. We then calculated the hourly wage given the maximum pay for hourly, monthly, and yearly pay periods. We assumed 160 hours for the monthly payments (40 hour work week for 4 weeks), and 2080 hours for the annual payments (40 hour work week for 52 weeks). Then, we saved the new data in a variable called `hourly_max_salary`.

Time Posted

Since the existing posted time is in different time zones, we converted the time format to EST time and only kept the hour. Then, we converted posted time from a quantitative variable to a categorical variable. Since it wouldn't make sense to have 24 different levels, we designated 4 levels: night (0 am - 5 am), morning (6 am - 11 am), afternoon (12 pm - 5 pm), and evening (6 pm - 11 pm).

Drop NA
We dropped all `NA` values for all predictors we wanted to observe so we can keep our dataset consistent when testing different models. Specifically, we dropped NAs for `hourly_max_salary`, `per_applies`, `follower_count`, `formatted_experience_level`, `original_listed_time`, and `remote_allowed`.

Filtering
To remove significant outliers that may affect the model's precision, we filtered and only kept the middle 90% of the data for the `hourly_max_salary` and `follower_count` variables. To generalize our results to job listings with a reasonable number of views, we filtered out job listings with less than 5 views. Because there were so few listings posted in the morning and night, we dropped the levels `morning` and `night` in `time_posted`.

Normalize Response Variable
Because each job has been listed for varying time durations, and its view count is directly related to its application count, we decided to normalize the application count with the view count. To do so, we created a new variable `per_applies`, which divides `applications` by `views`. We now use `per_applies` as our response variable.

Random Sampling

To address concerns with independence of jobs posted within the same company, we took a random sample with a size of 1500.

Model Type
Because `per_applies` is a numerical variable, a linear regression model would be most appropriate to predict the number of applications per view. As we addressed in the introduction, a person takes into consider many factors when applying to a job, so our model takes into consideration multiple predictors, including the hour of day posted, company follower count, experience level, maximum salary, ability to work remote, and if benefits are listed.

We checked for interaction effects between if remote work is allowed and maximum hourly salary, and level of experience and maximum hourly salary. The best-fit lines for maximum hourly salary by various levels of experience had different slopes, indicating a potential interaction effect, while the best-fit lines by remote work were parallel, indicating no interaction effect.

We split the `linkedin` dataset into training and testing data, with 75% of the data in training and 25% in testing. We then used cross-fold validation with 12 folds on the training data set to find the mean summary statistics (AIC, BIC, Adjusted R^2) for each model and compared the different values to find the best possible model. This process was repeated for three models, one full model, one with statistically significant variables, and one with statistically significant variables and including the interaction effect. We set a seed of (2) when splitting and folding the data to ensure reproducibility.

term	estimate	std.error	statistic	p.value
(Intercept)	0.210	0.009	23.387	0.000
hourly_max_salary	0.000	0.000	-0.848	0.397
follower_count	0.019	0.006	3.166	0.002
remote_allowed_yes	0.028	0.009	3.101	0.002
formatted_experience_level_Director	-0.034	0.019	-1.786	0.074
formatted_experience_level_Entry.level	-0.004	0.012	-0.324	0.746
formatted_experience_level_Executive	-0.062	0.037	-1.655	0.098
formatted_experience_level_Internship	-0.014	0.054	-0.255	0.799
formatted_experience_level_Mid.Senior.level	-0.006	0.010	-0.590	0.555
if_benefits_listed	-0.013	0.007	-1.786	0.074
time_posted_evening	-0.004	0.008	-0.496	0.620

mean_adj_rsq	mean_aic	mean_bic
0.018	-1436.347	-1377.085

Full Model: The mean adjusted R^2 value for a model including all predictors is 0.018, and the AIC and BIC was -1436.347 and -1377.085, respectively. Using a significance level of $\alpha = 0.10$, `follower_count`, `remote_allowed`, `if_benefits`, if the job requires director level of experience, and if the job requires executive level of experience are the only statistically significant variables and levels, with p-values of 0.002, 0.002, 0.074, 0.074, and 0.098 respectively. We then created a reduced model using only these statistically significant predictors.

mean_adj_rsq	mean_aic	mean_bic
0.019	-1439.242	-1389.857

Reduced Model: The reduced model had adjusted R^2, AIC, and BIC of 0.019, -1439.242, and -1389.857, respectively. Since this reduced model with less predictors had a higher adjusted R^2 and lower AIC and BIC, we concluded that the reduced model with `follower_count`, `remote_allowed`, and `formatted_experience_level` as predictors is a better model for predicting the percentage of applicants than the full model.

mean_adj_rsq	mean_aic	mean_bic
0.014	-1428.808	-1354.73

Interaction Effects Model: Lastly, we included a model with the variables from the reduced model, plus `hourly_max_salary` to explore the potential interaction effect between `hourly_max_salary` and `formatted_experience_level`, which we identified earlier has a potential interaction term. The adjusted R^2, AIC, and BIC are 0.014, -1428.808, and -1354.73. The adjusted R^2 is lower, and the AIC and BIC are higher for the reduced model with the interaction effect vs those of the reduced model without the interaction effect, indicating that the reduced model without the interaction term is a stronger model.

Overall, by comparing the adjusted R^2, AIC, and BIC values of each of the models (with all predictors, statistically significant predictors, and statistically significant predictors with an interaction effect), we concluded that the reduced model with `follower_count`, `remote_allowed`, and `formatted_experience_level` as predictors works as the best model.

Results

term	estimate	std.error	statistic	p.value
(Intercept)	0.211	0.009	24.669	0.000
follower_count	0.019	0.006	3.097	0.002
remote_allowed_yes	0.027	0.009	2.994	0.003
formatted_experience_level_Director	-0.040	0.018	-2.227	0.026
formatted_experience_level_Entry.level	-0.004	0.012	-0.313	0.754
formatted_experience_level_Executive	-0.067	0.037	-1.809	0.071
formatted_experience_level_Internship	-0.014	0.054	-0.256	0.798
formatted_experience_level_Mid.Senior.level	-0.008	0.009	-0.883	0.378
if_benefits_listed	-0.013	0.007	-1.794	0.073

$per\_applies = 0.211 + 0.019(follower\_count(millions)) + 0.027(remote\_allowed\_yes) - 0.040(Director) - 0.004(Entry\_Level) - 0.067(Executive) - 0.014(Internship) - 0.008(Mid\_Senior) - 0.013(if_benefits_listed)$

Intercept: When a company has 33,8808 followers, the hourly maximum salary is $54.64, there is no remote work allowed, and the experience level required is associate, 21.1% of viewers applied to the position.

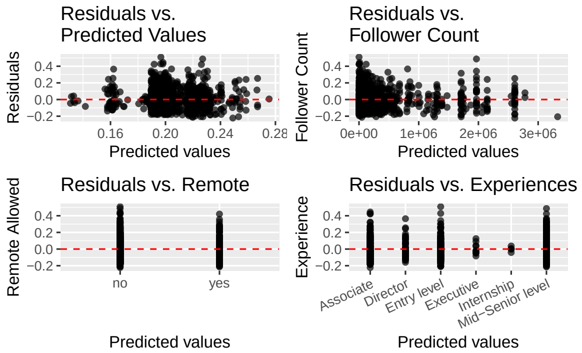
`Follower_count`: For every one million increase in followers a company has, the percentage of applicants from viewers increases by 1.9% on average, holding all else constant.

`Remote_allowed_yes`: We expect the percentage of applicants from viewers to be higher by 2.7%, on average, for positions that allow remote work vs positions that do not, holding all else constant.

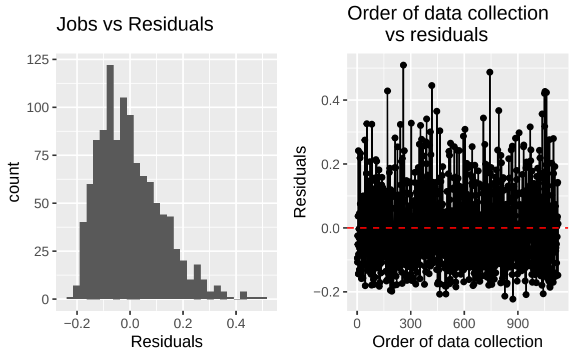
`Formatted_experience_level`: When the position requires the Director or Executive experience, the effect of formatted_experience_level is statistically significant ($\alpha = 0.10$), with p-values of 0.026 and 0.071, respectively. When the position requires director experience, we expect the percentage of applicants from viewers to be lower than that for requiring associate experience, on average, by 4.0%, holding all else constant. When the position requires executive experience, we expect the percentage of applicants from viewers to be lower than that for requiring associate experience, on average, by 6.7%, holding all else constant.

`if_benefits`: When the job lists benefits, we expect the percentage of applicants from viewers to be lower than that if benefits are not listed by 1.3% on average, holding all else constant.

Linearity: Since the plot of the residuals vs. predicted values do not have a discernible pattern and the plots of the residuals vs. each predictor do not have a discernible pattern, linearity is met.



Constant Variance: The vertical spread of the residuals is not constant in the plot of the residuals vs. predicted so the constant variance condition is not satisfied.



Normality: Based on "Jobs vs Residuals," the distribution of the residuals is unimodal but not symmetric, so the normality condition is not satisfied. However, the sample size is large enough to relax this condition since it is not satisfied.

Independence: Although there there were some jobs posted from the same company, which may influence independence within jobs, we took a random sample of the jobs to address this issue. Based on "Order of data collection vs residuals," there is clear pattern in the residuals vs. order of data collection plot, so independence condition appears to be satisfied, as far as we can evaluate it.

names	x
job_id	1.028
remote_allowedyes	1.010
formatted_experience_levelDirector	1.197
formatted_experience_levelEntry level	1.606
formatted_experience_levelExecutive	1.041
formatted_experience_levelInternship	1.021
formatted_experience_levelMid-Senior level	1.730
follower_count	1.017

Multicollinearity occurs where there are very high correlations among two or more predictor variables, and we need to check for multicollinearity because it causes a loss in precision in our estimates of the regression coefficients. All VIF values are less than 10, meaning multicolinearity is not a concern.

The RMSE of the training data is 0.1193, and the RMSE of the testing data is 0.1121. Significantly lower RMSEs for training data compared to the testing data could be a sign of model overfit, which means that the model fits the training data too well where it doesn't model new unseen data well. However, since the RMSE values for the training and testing data are very close, this shows no evidence of model overfit.

R Squared	Adj. R Squared	AIC	BIC
0.026	0.019	-1570.974	-1520.719

The low R^2 value of 0.023 suggests there is no significant relationship between the follower count, whether the job allows remote work, and the listed job experience level and the ratio of applications to views. Only 2.3% of the variation in the percent of viewers that applied is a result of the variation in the predictor variables: the follower count, whether the job allows remote, and the listed job experience level.

Discussion + Conclusion

Conclusion

Our final model suggests there is no statistically significant relationship between our response variable, the percent of viewers who applied, and the maximum salary, number of followers, having a remote option, the job experience level, the hour at which it was posted, and if the benefits were posted. Our model suggests the percentage of viewers who applied and the maximum salary, number of followers, having a remote option, the job experience level, the hour at which it was posted, and if the benefits were posted are not useful predictors, and therefore do not provide much insight into what makes a LinkedIn post have significant popularity.

Findings

Although our model is not a good fit, employers and employees can still take away two main findings. First, people's decisions about job applications are nuanced and difficult to predict using only measured metrics - factors outside of our model could contribute to employee application rates, such as company culture and personal interest could also contribute to view count and application rate.

Second, application rates could be impacted by applicants' belief about likelihood of success - sometimes viewer might not apply to avoid wasting their effort for a job they believe they are unqualified for. This is demonstrated in our model: variables such as `Director` and `Executive` have coefficients that are more negative, which means lower percentages of viewers applied. Consequently, variables that we hypothesized to be correlated with higher application rates, such as `hourly_max_salary` and `if_benefits`, have coefficients of approximately 0.000 and -0.007, respectively. This means they do not increase application rates among viewers, potentially because they are associated with jobs requiring more experience.

Limitations

Some of the limitations of our analysis may be our treatment of `NA` data in creating `if_benefits` and `remote_allowed`. We assumed that since the data had not been included, it was not available in the post, and therefore the values could be inputted as 0. Another possible limitation is that we chose to normalize the maximum salaries across hourly, monthly, and yearly pay periods to create `hourly_max_salary`, and to do so needed to make assumptions about how often individuals would work. These assumptions could have been incorrect and thus impacted the accuracy of our final model.

Additionally, some limitations are due to our dataset and scraped data. Other possible influences to the percentage of applicants that were not scrapped could be the number of shares the post received, the duration of the post, and if the location as if it was rural or not. There are also other less quantifiable variables that go into applying to a job, such as company culture and the company's mission statement, as well as societal influences such as current unemployment rates. To improve the prediction of applications per view, we might have to consider a wider range of variables that were not in the data sets we used.

Future Improvements

To improve our analysis, we hope to obtain industry information and create data subsets since application patterns might differ significantly across industries - some might have faster turnover rates depending on economic trends or have higher demand for different levels of positions at different seasons, such as summer internships. Additionally, we hope to fix issues regarding time scraped from LinkedIn, since many of the original listed times are the exact same as the listed times (the time it was scraped), which is not meaningful in determining application rates over time. Lastly, we hope to potentially fill in `NA` values since there could be a skew in the remaining data that influenced our results.