Regression Analysis

Predicting Company Popularity & Employee Satisfaction

Matias Beeck

The Questions

- How does the <u>amount of maternity leave</u> offered by a company affect it's reputation, popularity or employee satisfaction?

- What other company data could play into employee satisfaction?

		Maternity Leave		*Paternity Leave*		
Name	**Industry**	**Paid (weeks)**	**Unpaid (weeks)**	**Paid (weeks)**	**Unpaid (weeks)**	**Add a Tip**
Netflix	Technology: Consumer Internet	52	0	52	0	
Bill and Melinda Gates Foundation	Philanthropy	52	N/A	52	2	
Army (British)	Government: Federal	39	13	2	N/A	
Automattic, Inc.	Technology: Consumer Internet	32	0	N/A	N/A	
Ford Motor	Automotive: Manufacturers	30	4	0	N/A	
Zurich	Insurance: Life	29	N/A	0	N/A	
Etsy	Technology: Consumer Internet	26	N/A	26	N/A	

*Fairygodboss.com crowd sourced maternity leave data

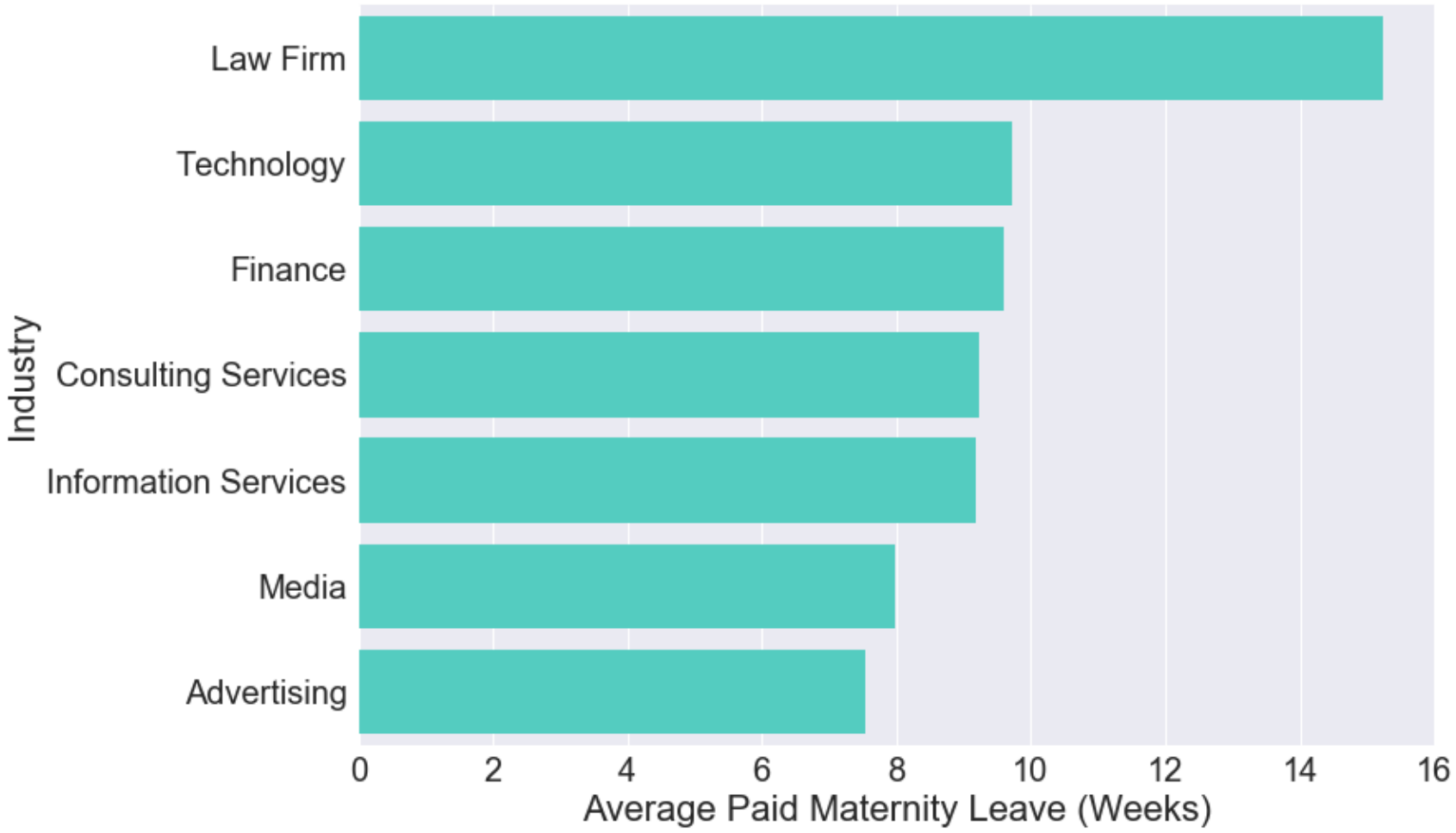
The Data

Scraping Company Data

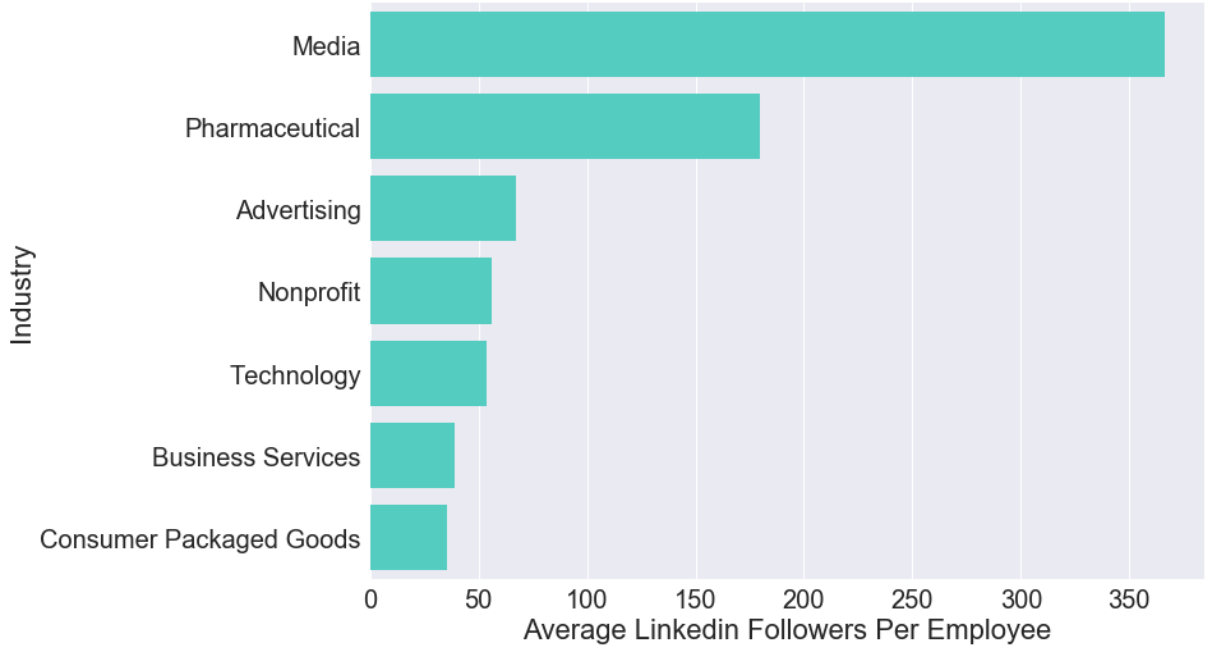
- **Fairygodboss.com dataset:**
  - Maternity leave info for ~1700 companies
  - Paid Maternity Leave (weeks)
  - Unpaid Maternity Leave (weeks)
  - Paid Paternity Leave (weeks)
  - Unpaid Paternity Leave (weeks)
  - Industry
- **Linkedin Followers**
- **HQ Population data**

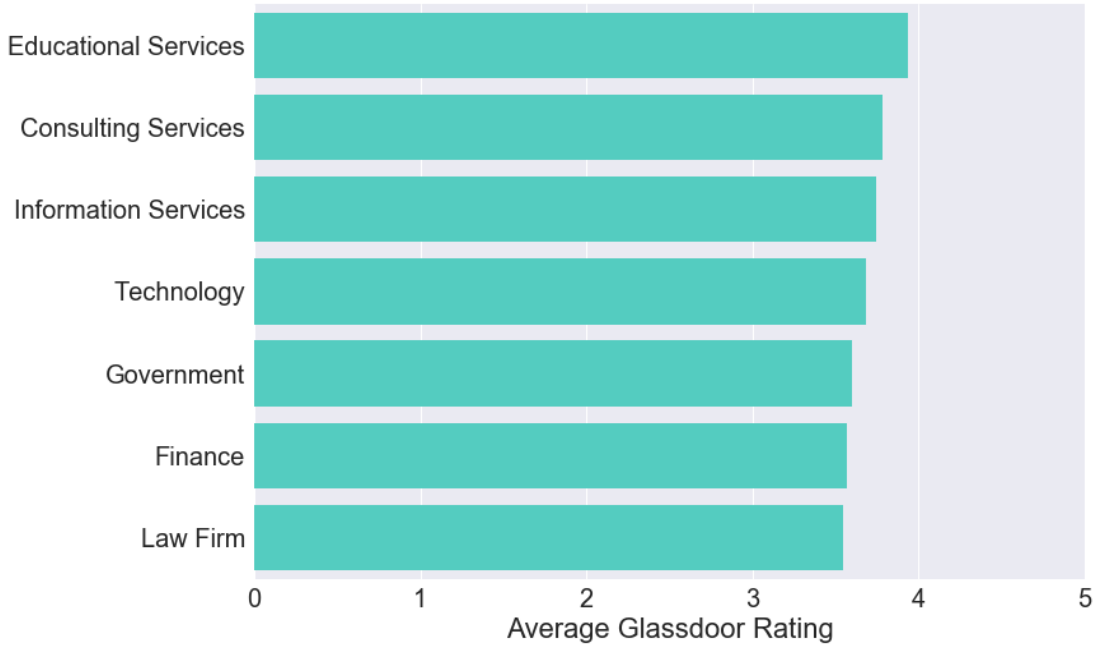
- **Glassdoor data:**
  - Company Glassdoor rating (1-5)
  - Revenue per year
  - Employee number
  - # of employee reviews
  - CEO approval %
  - Interview difficulty
  - Interview experience (% of positive, neutral or negative)
  - Year Founded (age)
  - Headquarter Location

Exploring Dataset



Exploring Dataset: Dependent Variables





Linear Regression

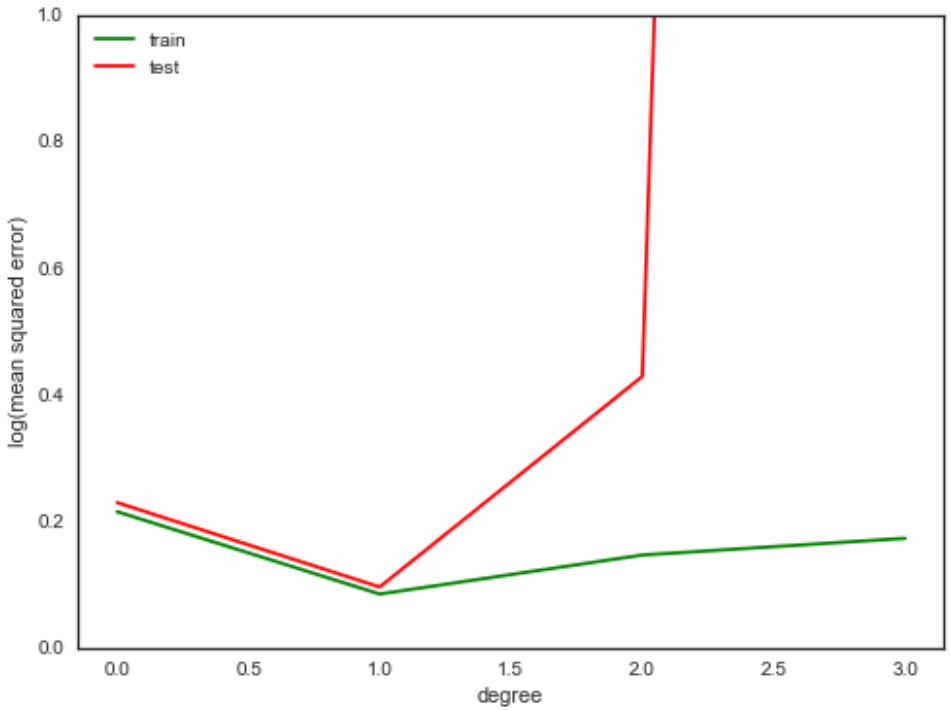
Linear Regression: Kitchen Sink

1. Predicting "popularity" of companies
   - Dependent variable = Log of Linkedin followers per employee
   - Adjusted **R^2 = .174**

2. Predicting employee satisfaction
   - Dependent variable = Glassdoor rating (1-5)
   - What factors play into employees liking a company?
   - Adjusted **R^2 = .611**

Linear Regression: Predicting GD Rating

- Using all features
- Adjusted R^2 = .611
- Polynomial analysis using test, train ->
- Cross Validation (K-fold)
  - 5 folds = .609 R^2
  - 10 folds = .601 R^2



Linear Regression: "Improved" Model

- Removed:
  - Industry dummy variables
  - Number of employee reviews
  - HQ population
- Transformed:
  - Log of employee count
- Adjusted R^2 = .545
- Cross Validation (K-fold)
  - 5 folds = .545 R^2
  - 10 folds = .54 R^2

Dep. Variable:	gd_rating	**R-squared:**	0.549
Model:	OLS	**Adj. R-squared:**	0.545

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.5964	0.123	12.980	0.000	1.355	1.838
mat_paid_weeks	0.0043	0.002	2.344	0.019	0.001	0.008
gd_ceo_approval	0.0159	0.001	26.398	0.000	0.015	0.017
co_age	0.0004	0.000	2.092	0.037	2.62e-05	0.001
log_linkedin_followers	0.0342	0.006	5.916	0.000	0.023	0.046
gd_interview_pos_per	0.0056	0.001	7.045	0.000	0.004	0.007
gd_interview_difficulty	0.1391	0.031	4.476	0.000	0.078	0.200
revenue	-6.81e-12	3.35e-12	-2.032	0.042	-1.34e-11	-2.34e-13
log_employee_num	-0.0524	0.010	-5.105	0.000	-0.073	-0.032

Linear Regression: Interpretation

- <u>Predicting Glassdoor Rating</u>
- Surprising:
  - Variation in paid maternity leave, company age & revenue explained little of variation in GD rating
  - Interview stats were somewhat important
- Most predictive power in model from CEO approval rate
  - Adjusted R^2 drops to .225 without it

Dep. Variable:	gd_rating	**R-squared:**	0.549
Model:	OLS	**Adj. R-squared:**	0.545

feature	abs_coefficient
gd_ceo_approval	0.60
gd_interview_pos_per	0.15
log_employee_num	0.14
log_linkedin_followers	0.13
gd_interview_difficulty	0.10
mat_paid_weeks	0.05
revenue	0.05
co_age	0.04

Linear Regression

Tech Industry Focus

Linear Regression: Tech Industry

- Removed:
  - Company age
  - Revenue
  - Linkedin followers
- Adjusted R^2 = .74
- Cross Validation (K-fold)
  - 5 folds = .73 R^2
  - 10 folds = .72 R^2
- Most predictive power comes from CEO approval rate
  - Adjusted R^2 drops to .323 without it

Dep. Variable:	gd_rating	**R-squared:**	0.738
Model:	OLS	**Adj. R-squared:**	0.728

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	0.9682	0.292	3.318	0.001	0.391	1.546
mat_paid_weeks	0.0074	0.003	2.202	0.029	0.001	0.014
gd_ceo_approval	0.0214	0.002	14.187	0.000	0.018	0.024
gd_interview_pos_per	0.0045	0.002	2.350	0.020	0.001	0.008
gd_interview_difficulty	0.4997	0.091	5.511	0.000	0.320	0.679
log_employee_num	-0.1023	0.016	-6.253	0.000	-0.135	-0.070

Linear Regression: Tech Industry Lasso Coefficients

feature	abs_coefficient
gd_ceo_approval	0.70
log_employee_num	0.31
gd_interview_difficulty	0.26
gd_interview_pos_per	0.13
mat_paid_weeks	0.11

Dep. Variable:	gd_rating	**R-squared:**	0.738
Model:	OLS	**Adj. R-squared:**	0.728

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.9682	0.292	3.318	0.001	0.391	1.546
mat_paid_weeks	0.0074	0.003	2.202	0.029	0.001	0.014
gd_ceo_approval	0.0214	0.002	14.187	0.000	0.018	0.024
gd_interview_pos_per	0.0045	0.002	2.350	0.020	0.001	0.008
gd_interview_difficulty	0.4997	0.091	5.511	0.000	0.320	0.679
log_employee_num	-0.1023	0.016	-6.253	0.000	-0.135	-0.070

Next Steps

Challenges & Next Steps

- Explore the tech industry subset for potential interaction features
- Gather more companies for dataset
  - Eliminate maternity leave data altogether
  - Companies in dataset could be highly bias

Thanks!