Big Data Salary Prediction

With Linear Regression

Yuwen Huang

Introduction

Objection

- Predicting annual salary for big data career by data collected from employment website.

Goal- Create the optimal model to predict salary

- Help people in job industry to find expected salary before negotiating.
- For people who already in data related industry, they can use this model to check if they get reasonable salary

Data Processed

Target Variable:

Annual Salary

Web Scraping tool:
Beautifulsoup & Selenium
Website:
indeed

Features Variables:

About Company:

- Headquarter_State
- Company Annual Revenue
- Company Size
- Founded year
- Industry

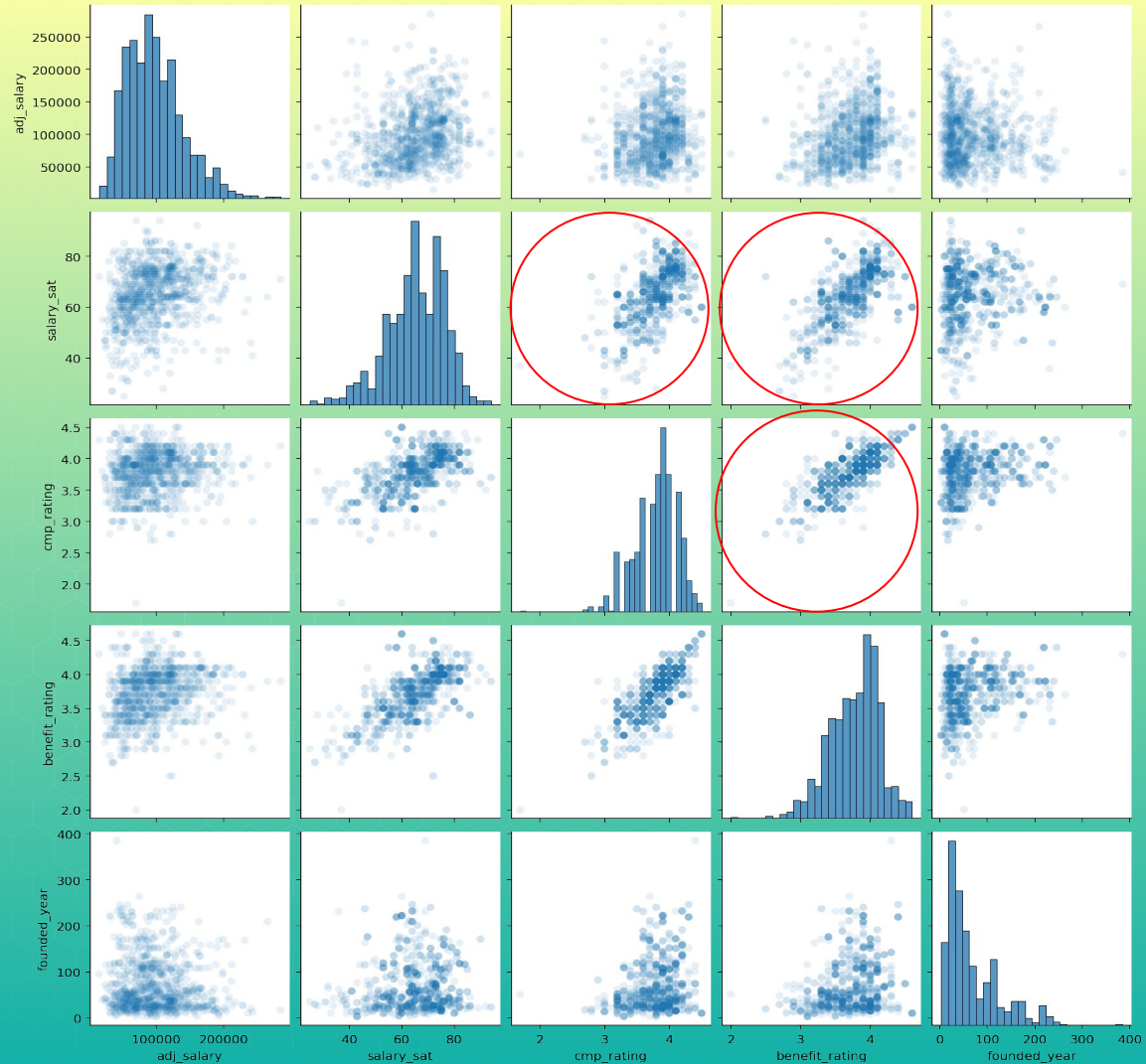
Company Review:

- Salary Satisfaction %
- Company Overall Rating
- Company Benefit Rating

Job:

- Job title
- Interview Process Length
- Interview Difficulty
- Payment Type

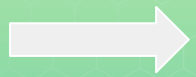
Pairs Plot for Numerical Variables



Preliminary Linear Regression

Linear Regression with all features variables

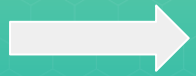
R^2 = 0.5184
R^2 = 0.4831
R^2 = 0.5603



Avg R^2 = **0.5206**

Polynomial Regression with all features variables

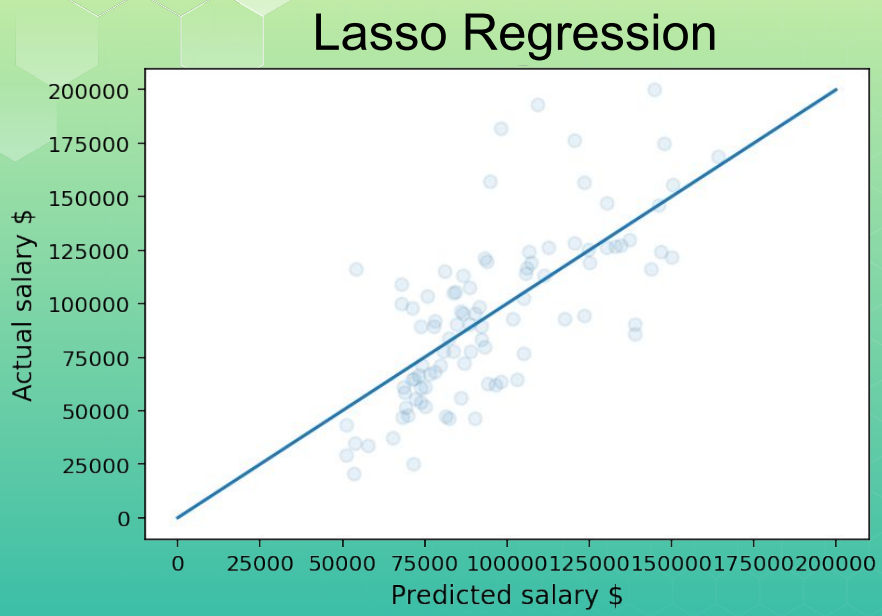
R^2 = 0.5252
R^2 = 0.4886
R^2 = 0.5699



Avg R^2 = **0.5279**

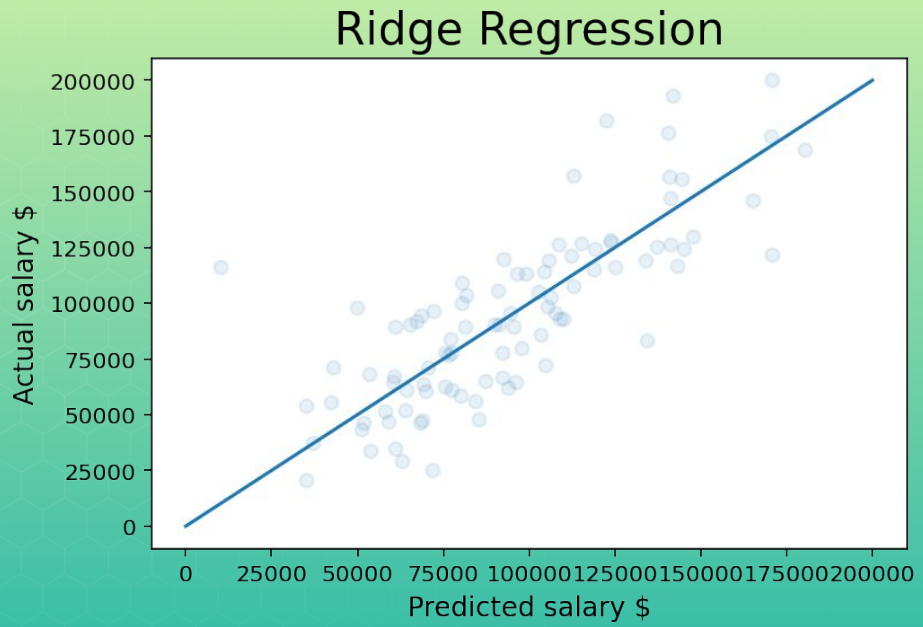
LASSO Regression VS. Ridge Regression

Regularization



R^2 = 0.51

MAE= 22,105.59



R^2 = 0.61

MAE= 18,331.80

LASSO Regression Variable Selection

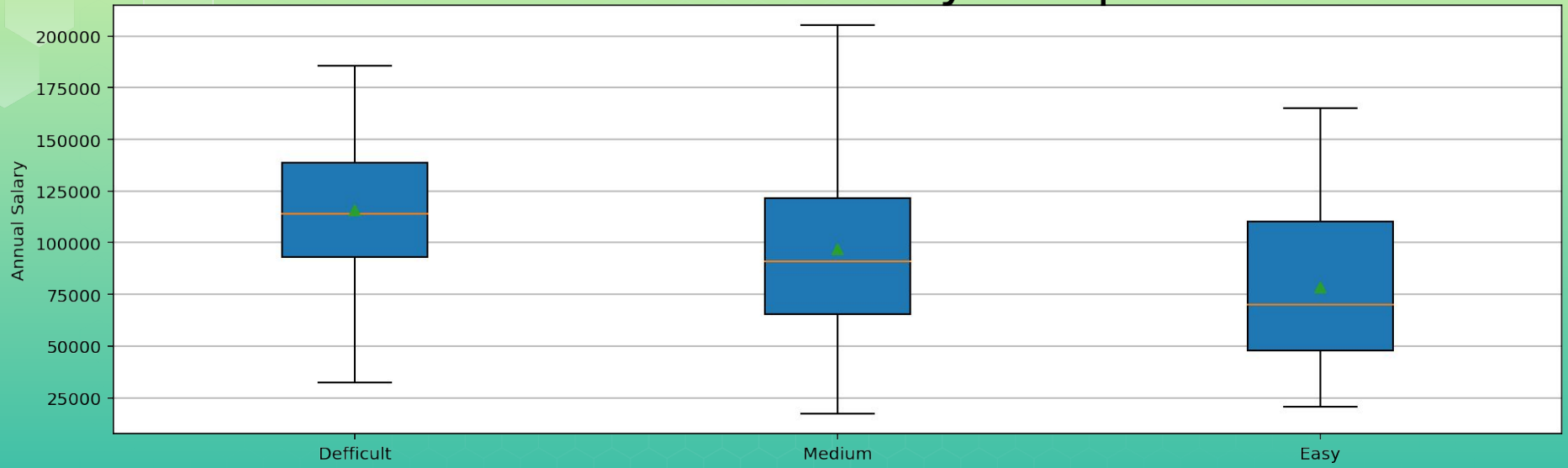
Variables	Coefficient
title.analyst salary_sat	-411.457
title.analyst founded_year	-16.27
title.senior salary_sat	157.12
title.senior founded_year	3.40
title.scientist salary_sat	354.76
Headquarters_State_DC founded_year	-1.33
Headquarters_State_Other salary_sat	-10.44
revenue_$1B to $5B (USD) founded_year	59.50
revenue_more than $10B (USD) salary_sat	7.82

Variables	Coefficient
inter_len_About a month founded_year	9.84
inter_len_About two weeks salary_sat	63.99
inter_len_More than one month founded_year	-25.67
salary_unit_per year salary_sat	400.26
Energy, Mining & Utilities founded_year	13.75
Government & Public Administration salary_sat	-8.95
salary_sat^2	5.243
salary_sat founded_year	-0.49
founded_year^2	-0.02

Remove Non-Statistically Significant Variables to optimal Ridge Model

Removed Variable	Current R^2	Updated R^2	R^2 Improvement
Company Size	0.61	0.64	0.03
Company Overall rating	0.64	0.65	0.01
Benefit Rating	0.65	0.66	0.01
Interview Difficulty	0.66	0.55	**-0.11**

Interview Difficulty box plot



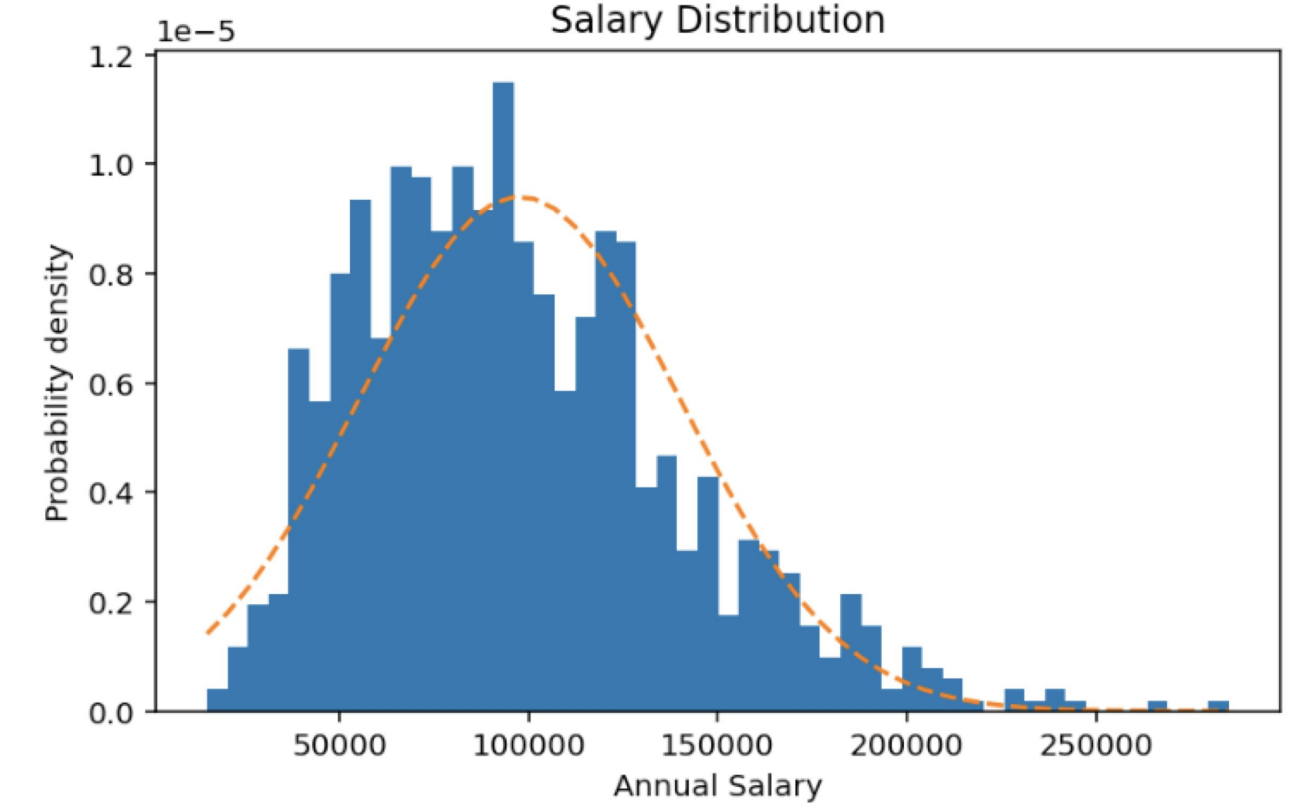
Final Model & Conclusions

- **The optimal model:**
  - Ridge Regression Model
- **Including features:**
  - Headquarter_State
  - Company Annual Revenue
  - Founded year
  - Salary Satisfaction %
  - Job title (scientist, analyst, senior)
  - Interview Process Length
  - Interview Difficulty
  - Payment Type
- **R^2:** 0.666
- **Mean Absolute Error:** 16831.94

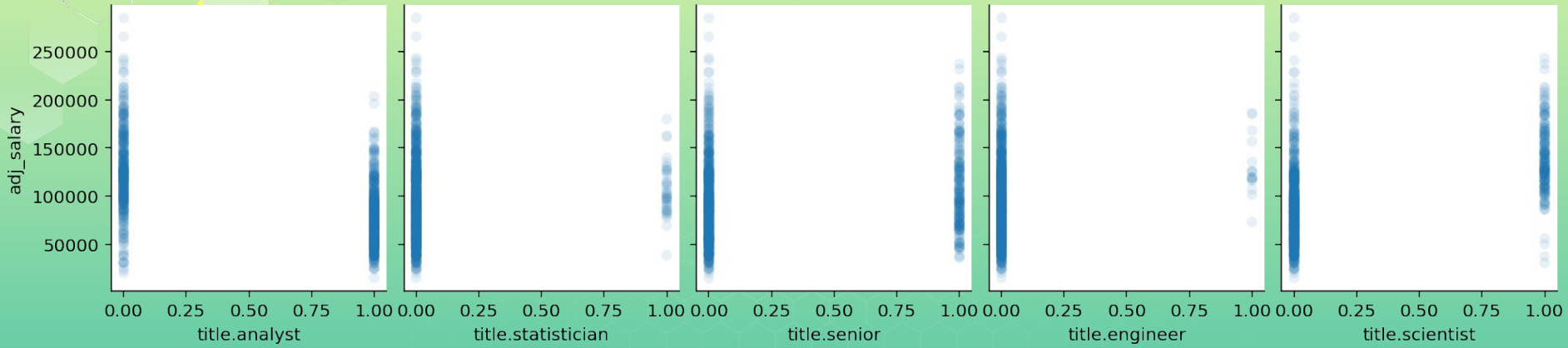
THANK YOU

APPENDIX

Distribution of Annual Salary

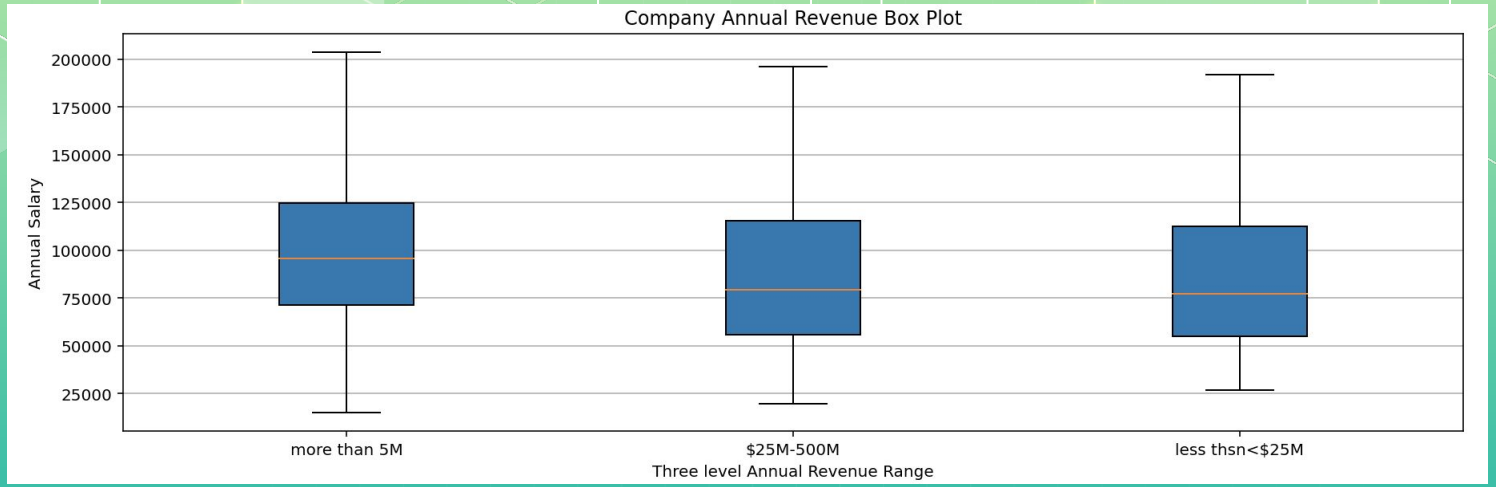


Pairs Plot for Title Variables



Remove title.statistician and title.engineer from ridge model:
R^2 score increase 0.004 to 0.666.

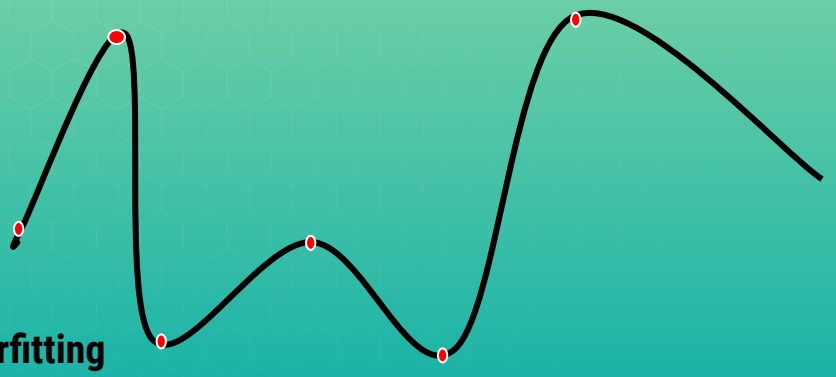
BoxPlot of Company Annual Revenue



Polynomial Regression

R^2 for training data is **0.916**

R^2 for testing data is **- 3.3**



Overfitting

('title.senior', -5762.426346233193),
('title.scientist', 985.8822957477951),
('Headquarters_State_CT', 6188.745788723318),
('Headquarters_State_DC', 202.08306782503792),
('Headquarters_State_FL', 1117.9844778638997),
('Headquarters_State_GA', -220.64875407813315),
('Headquarters_State_IL', -921.3283836144151),
('Headquarters_State_MA', -754.7972169554096),
('Headquarters_State_MD', -526.2022470509401),
('Headquarters_State_MN', 641.3332858633435),
('Headquarters_State_NC', -470.1012289344508),
('Headquarters_State_NJ', 682.1501291217144),
('Headquarters_State_NY', -818.0260283338703),
('Headquarters_State_Other', -251.65588938198744),
('Headquarters_State_TN', 167.62429886666996),
('Headquarters_State_TX', -30.94786531332454),
('Headquarters_State_VA', 428.4371344457495),
('Headquarters_State_WA', 1849.0129016708506),
('Headquarters_State_WI', 770.5408505907744),
('revenue_$1B to $5B (USD)', 314.7564727284416),
('revenue_$1M to $5M (USD)', -3360.7248734854857),
('revenue_$25M to $100M (USD)', -27.73111185644302),
('revenue_$500M to $1B (USD)', 619.3579974454169),
('revenue_$5B to $10B (USD)', 4664.067912251381),
('revenue_$5M to $25M (USD)', -1884.482517891447),
('revenue_less than $1M (USD)', 144.22199985563157),
('revenue_more than $10B (USD)',(-460.11154569072164)

('inter_diff_Easy', 322.4279199996231),
('inter_diff_Medium', -1078.3239247232077),
('inter_diff_none', 1735.9507493741849),
('inter_len_About a month', -712.5190726615895),
('inter_len_About a week', 1977.6141850854697),
('inter_len_About two weeks', 656.94997482089),
('inter_len_More than one month', 1168.941275478176),
('inter_len_none', -1703.1611623214235),
('salary_unit_per month', 1644.850613584671),
('salary_unit_per year', -230.91775638753856),
('Aerospace & Defense', 2268.9577127856664),
('Education', -183.2852442346466),
('Energy, Mining & Utilities', 19.364111537963936),
('Financial Services', -42.16054278952453),
('Government & Public Administration', 332.03296850658626),
('Healthcare', 260.42133738771645),
('Human Resources & Staffing', -1488.6333514954922),
('Information Technology', 591.7546380215108),
('Insurance', -843.4599369498455),
('Manufacturing', 2164.3388443918657),
('Pharmaceutical & Biotechnology', 23.213216077430843),
('Retail & Wholesale', -1467.122205810767),
('Telecommunications', -697.4622294364681),
('salary_sat', -445.44321887897524),
('founded_year', -136.9720554654632)

Further Improvements

- Add more feature variables: work experience, skill requirement
- Collect more observations