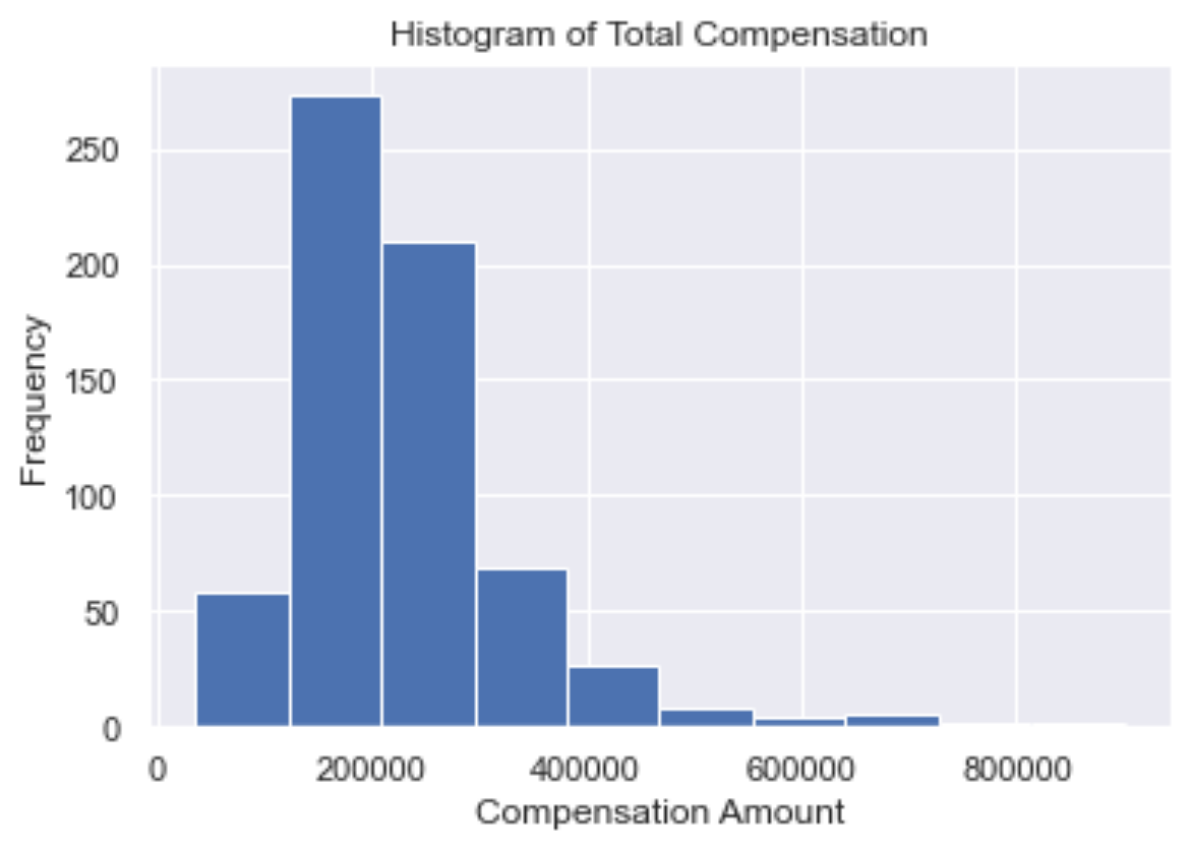

SALARY
PREDICTION
A Linear Regression Analysis

Can we predict a data scientist's compensation?

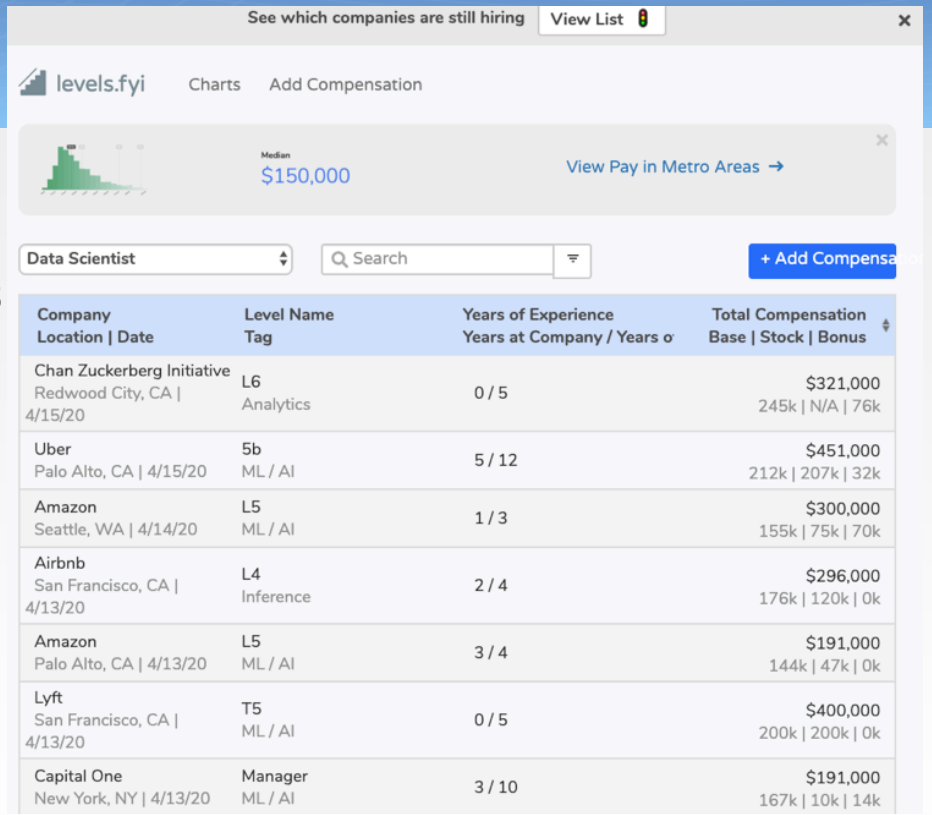
Why ask the question?

As an incoming data scientist, I am curious to know what the data reveals…



Methodology

- **Data sources:**
  - levels.fyi 700+ data scientist entries
  - salary.com cost of living data
  - Wikipedia The Free Encyclopedia company information
- **Model:**
  - **Ridge regression**
  - alpha = 150
  - scaled features



A First Look at the Data…

Total Compensation
average:

~$229,000
(*s* = $107,000)

Years at Company
average:

2 years
(*s* = 2.5 years)

Years of Experience
average:

5 years
(*s* = 4.4 years)

Ridge Regression Results

alpha = 150; scaled features

R^2	MAE
0.236*	$64,014

* up from a low point of R^2 = 0.087!

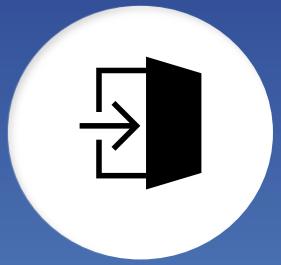
Interpretations

Interesting Insights

Years of Experience

Product & ML / AI

Company Age



Years at Company

↑ Years of Work Experience

- most influential feature
- 1 additional year corresponds to ~$11K more in compensation

↓ Company Age

- negative coefficient suggesting potentially less pay at a mature company

↑ Product & Machine Learning / AI

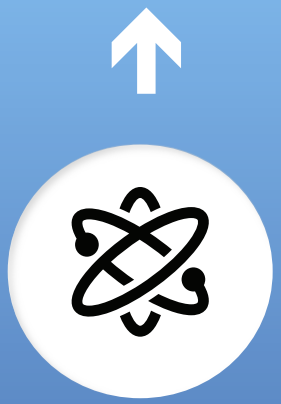
- focus areas with greatest coefficient results, indicating higher compensation relative to other areas

↓ Years at Company

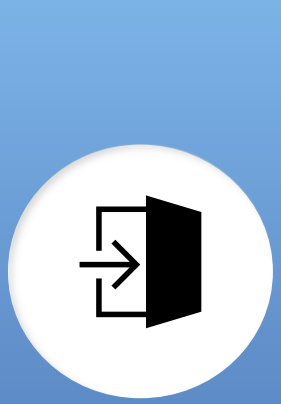
- negative coefficient
- 1 additional year corresponds to ~$4.5K less in compensation

Influence on Compensation

Years of Experience



Product & ML / AI



Years at Company

Company Age

Future Work

- Masters / PhD feature
- Company revenue, net income
- Base – stock – bonus breakdown
- Get more compensation data!

In Conclusion…


TAKE THE MONEY AND RUN!
memegenerator.net

Appendix

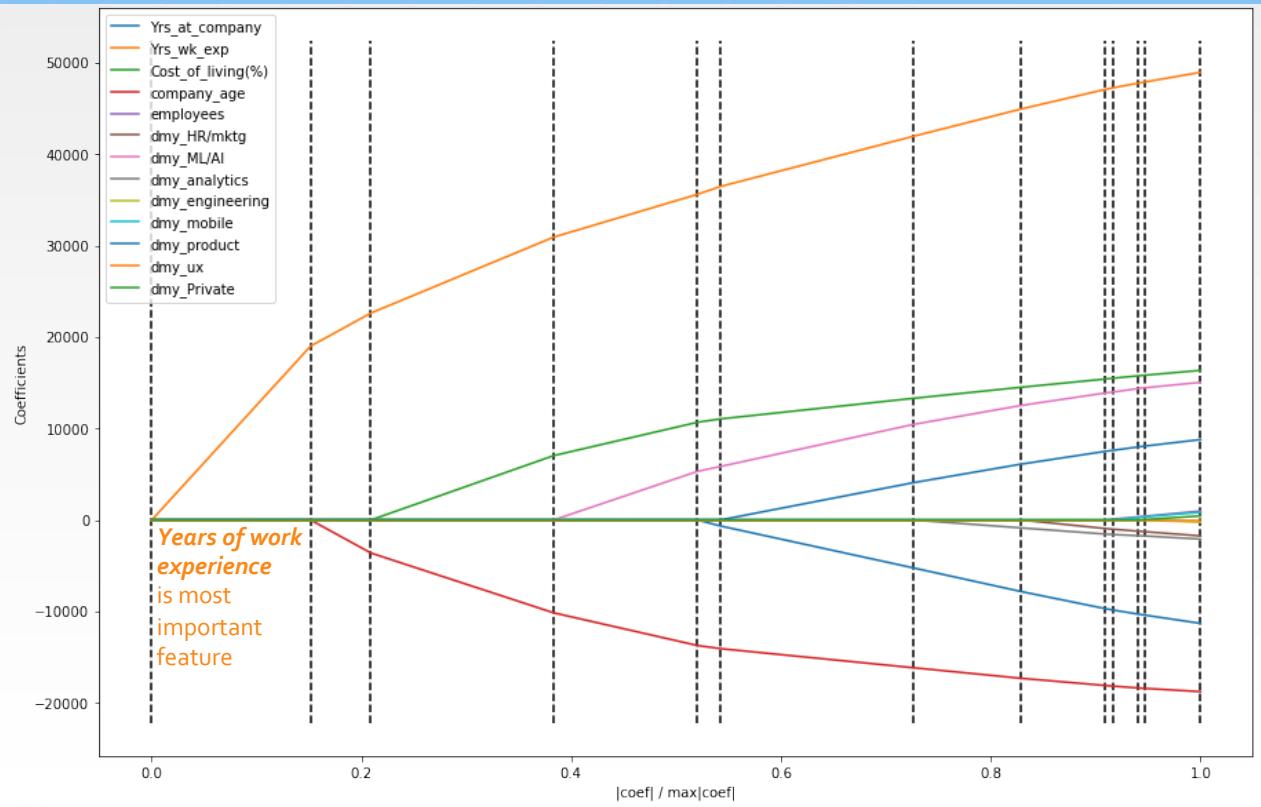
Ridge Model Coefficients

Note: ridge model coefficients are scaled; to return scaled coefficients to interpretable figures, divide by standard deviations

	scaled coefs:	*std:*	*unscaled coefs:*
Years at company	-11,196.81	2.48	-4,514.84
Years of work experience	48,758.44	4.35	11,208.84
Cost of living (%)	16,315.26	24.32	670.86
Company age	-18,726.30	29.33	-638.47
Employees at company	971.08	357,586.46	0.00
Private company	457.11	0.24	1,904.64

focus areas:	*scaled coefs:*	*std:*	*unscaled coefs:*
HR / marketing	-1,740.18	0.07	-24,859.71
Machine learning / AI	14,942.73	0.47	31,793.04
Analytics	-2,117.41	0.26	-8,143.88
Engineering	-273.95	0.25	-1,095.80
Mobile	833.2	0.13	6,409.23
Product	8,732.12	0.20	43,660.60
UX	0.00	0.06	0.00

LASSO Regularization Path using LARS



- LARS Path is a tool for visualizing feature importance
- At far left is value of alpha for which the penalty on coefficients is largest (coefs are 'zeroed out')
- At far right, there is no penalty, equivalent to a 'regular' linear regression
- **Features that enter model earliest (from the left) are ones the model treats as most essential**