Web Scraping: Data Science Job Postings

Mark Mummert
April 4, 2017

Objective

Find the most important features in job postings that predict whether the pay will be above or below the median

Outline

- Gathering Data
  - Webscraping
  - Data
- Processing
  - Feature selection
- Final Results
- Discussion

Collecting Data

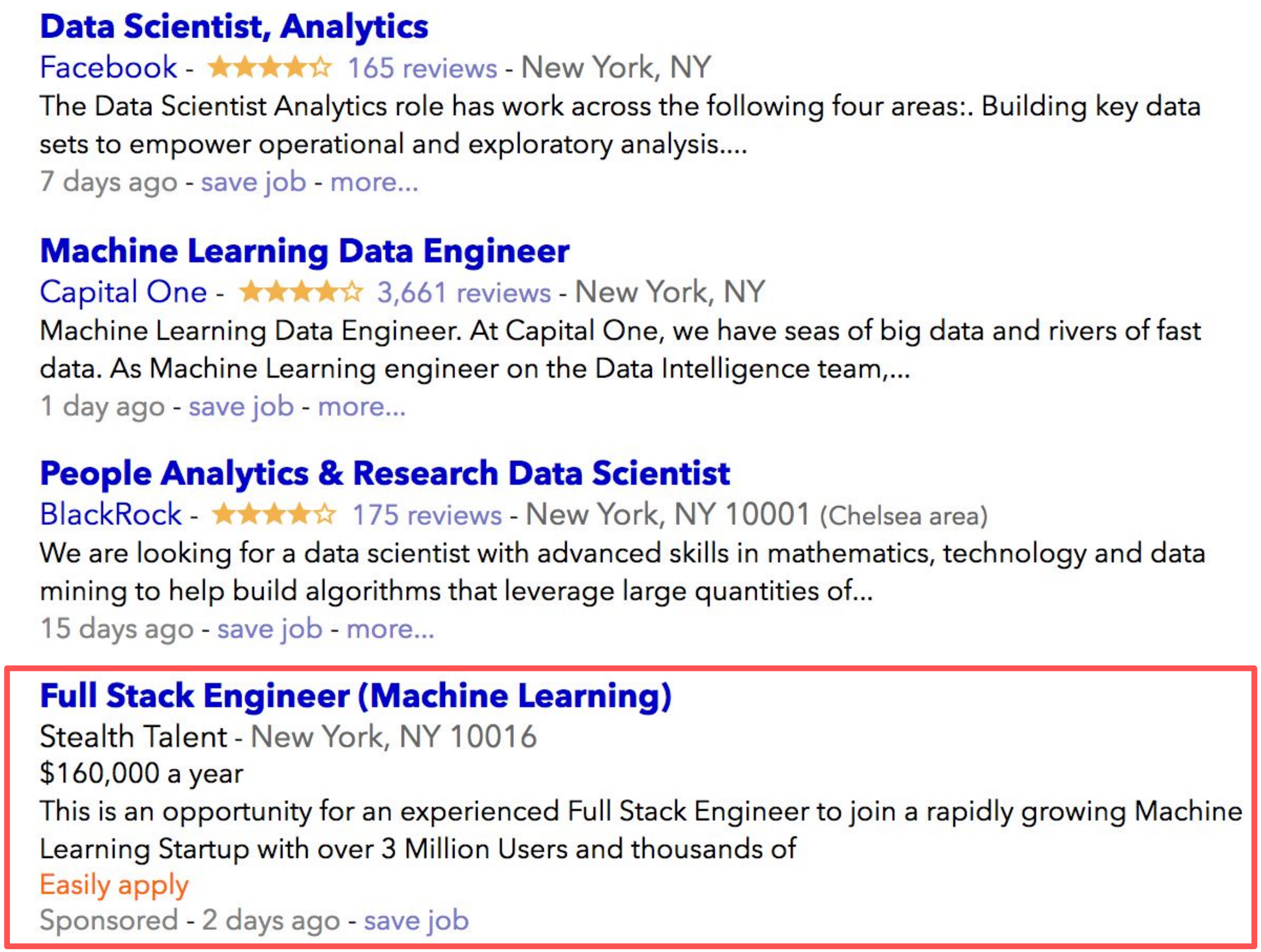
‘Scraping’ Indeed.com

- Search results for Data Science
- April 9 - 11
- 3-4 times per day

New York
Chicago
San Francisco
Austin
Seattle

Los Angeles
Philadelphia
Atlanta
Dallas
Pittsburgh
Portland

Phoenix
Denver
Houston
Miami
Washington



Data Scientist, Analytics

Facebook - ★★★★☆ 165 reviews - New York, NY

The Data Scientist Analytics role has work across the following four areas:. Building key data sets to empower operational and exploratory analysis....

7 days ago - save job - more...

Machine Learning Data Engineer

Capital One - ★★★★☆ 3,661 reviews - New York, NY

Machine Learning Data Engineer. At Capital One, we have seas of big data and rivers of fast data. As Machine Learning engineer on the Data Intelligence team,...

1 day ago - save job - more...

People Analytics & Research Data Scientist

BlackRock - ★★★★☆ 175 reviews - New York, NY 10001 (Chelsea area)

We are looking for a data scientist with advanced skills in mathematics, technology and data mining to help build algorithms that leverage large quantities of...

15 days ago - save job - more...

Full Stack Engineer (Machine Learning)

Stealth Talent - New York, NY 10016

$160,000 a year

This is an opportunity for an experienced Full Stack Engineer to join a rapidly growing Machine Learning Startup with over 3 Million Users and thousands of

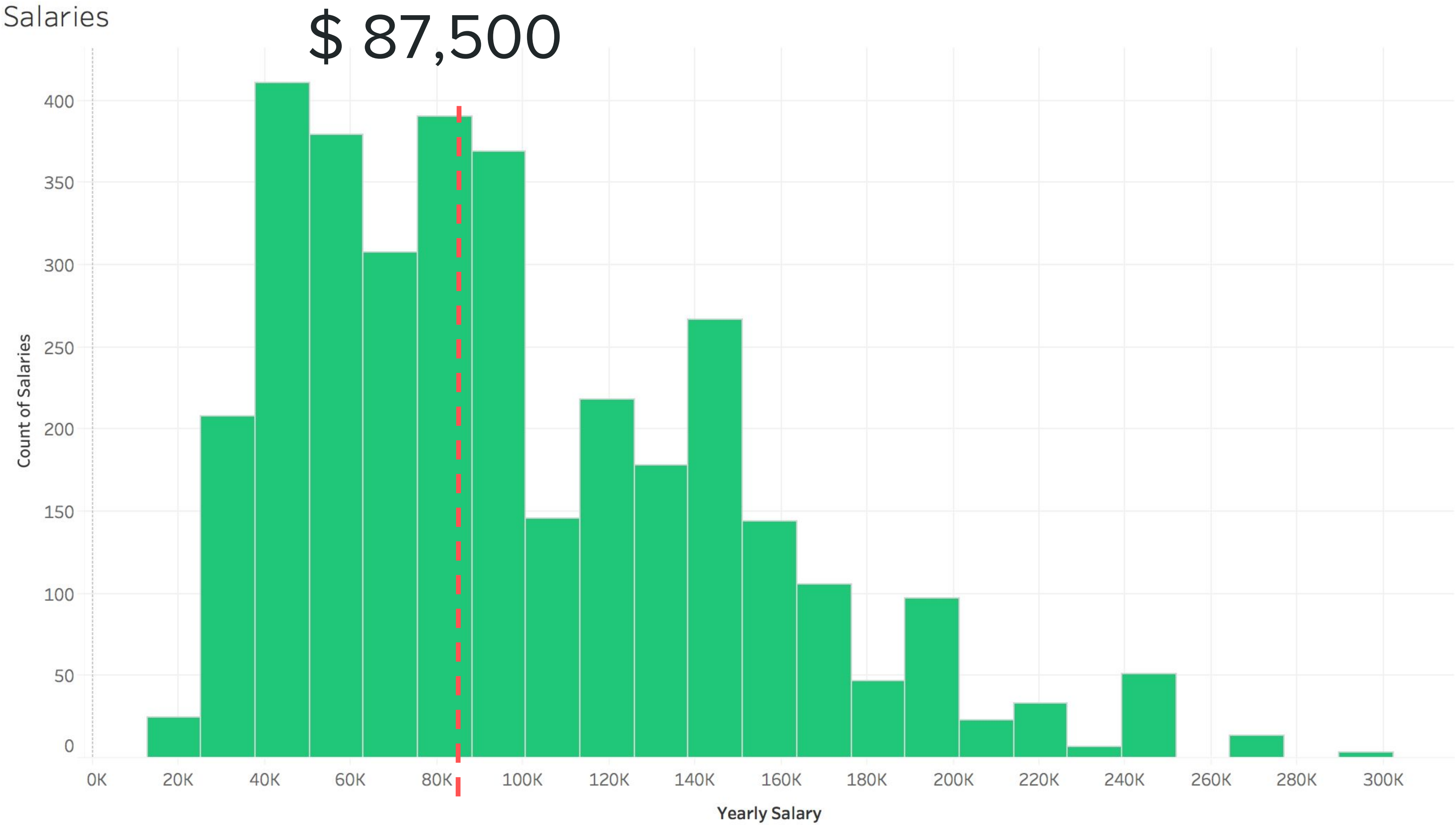
Easily apply

Sponsored - 2 days ago - save job

Collection Results

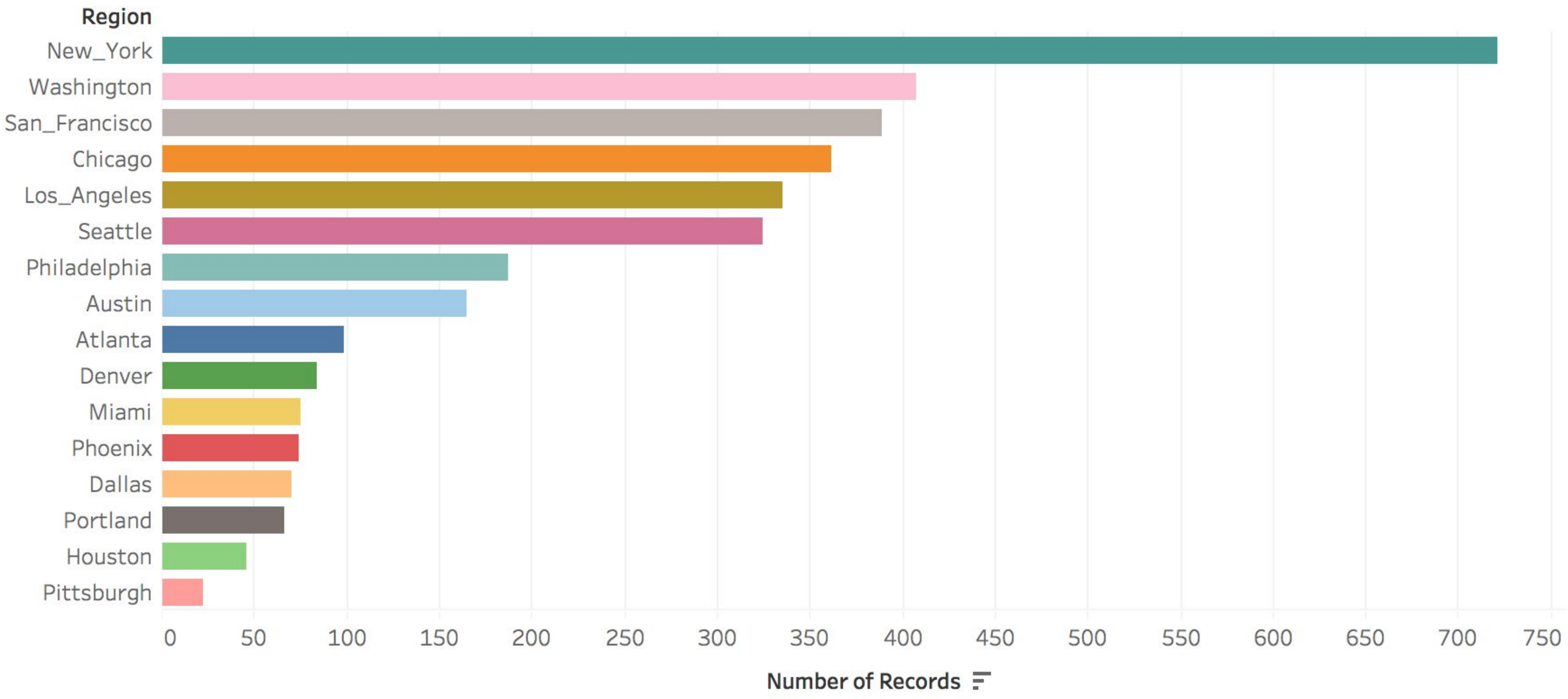
Results: 58,684

Salaries: 3,424



Salaries

$ 87,500


Region
New_York
Washington
San_Francisco
Chicago
Los_Angeles
Seattle
Philadelphia
Austin
Atlanta
Denver
Miami
Phoenix
Dallas
Portland
Houston
Pittsburgh
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
Number of Records

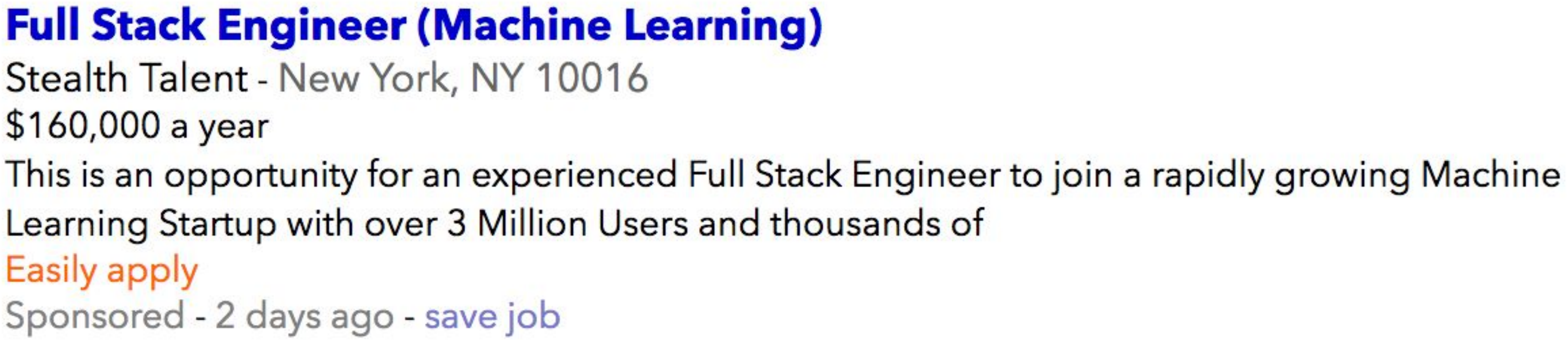
Language Processing

What is language processing?

A system to convert words to numeric features

Used a binary system - does a word appear in a result or not?

Use a classifier to see how presence of those words impact salary



Full Stack Engineer (Machine Learning)

Stealth Talent - New York, NY 10016

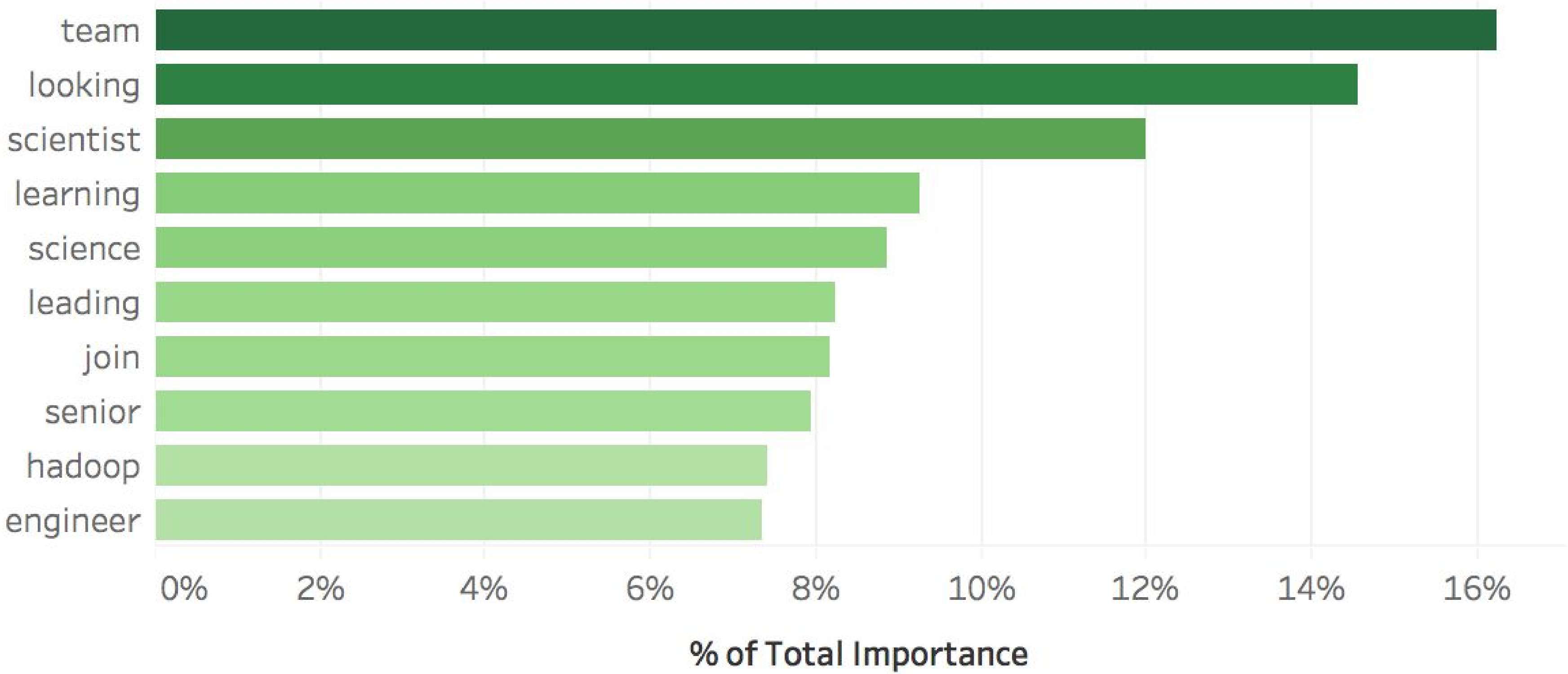
$160,000 a year

This is an opportunity for an experienced Full Stack Engineer to join a rapidly growing Machine Learning Startup with over 3 Million Users and thousands of

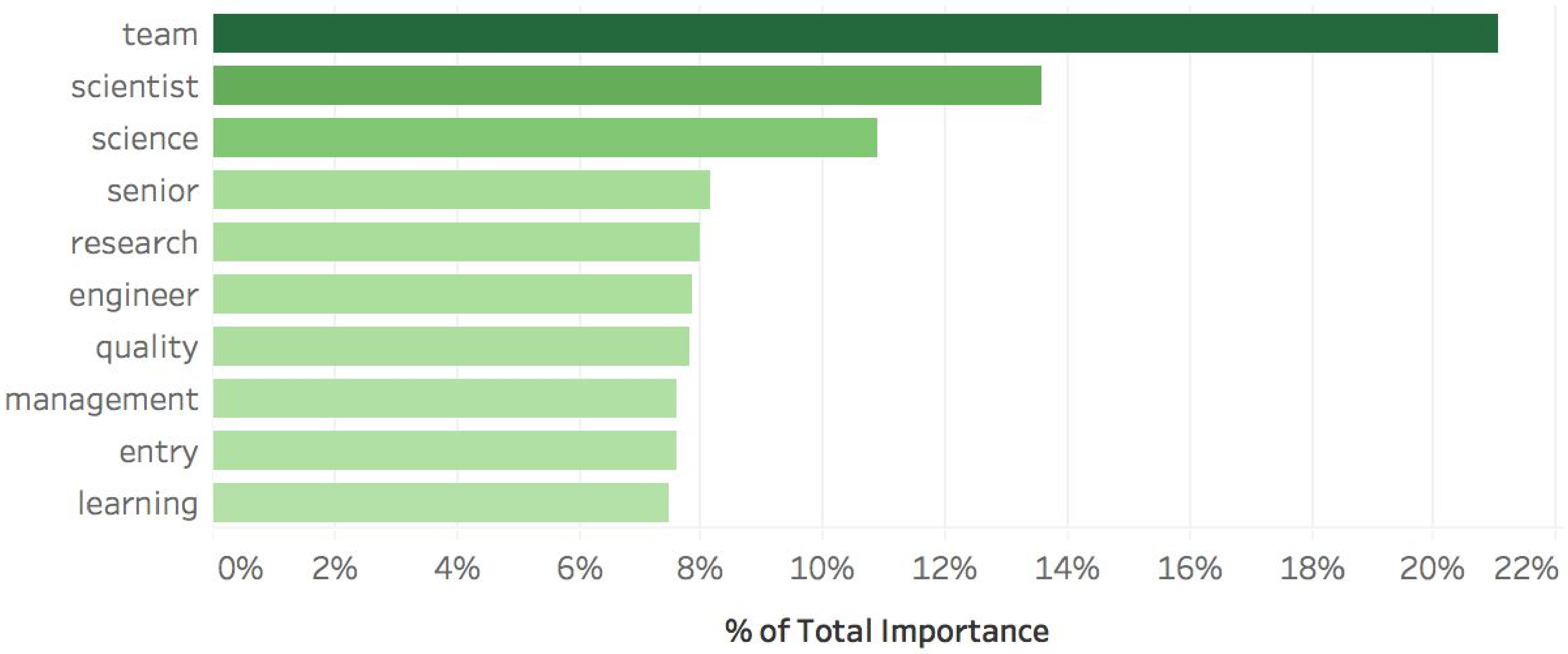
Easily apply

Sponsored - 2 days ago - save job

Most Relevant Description Words



Most Relevant Title Words

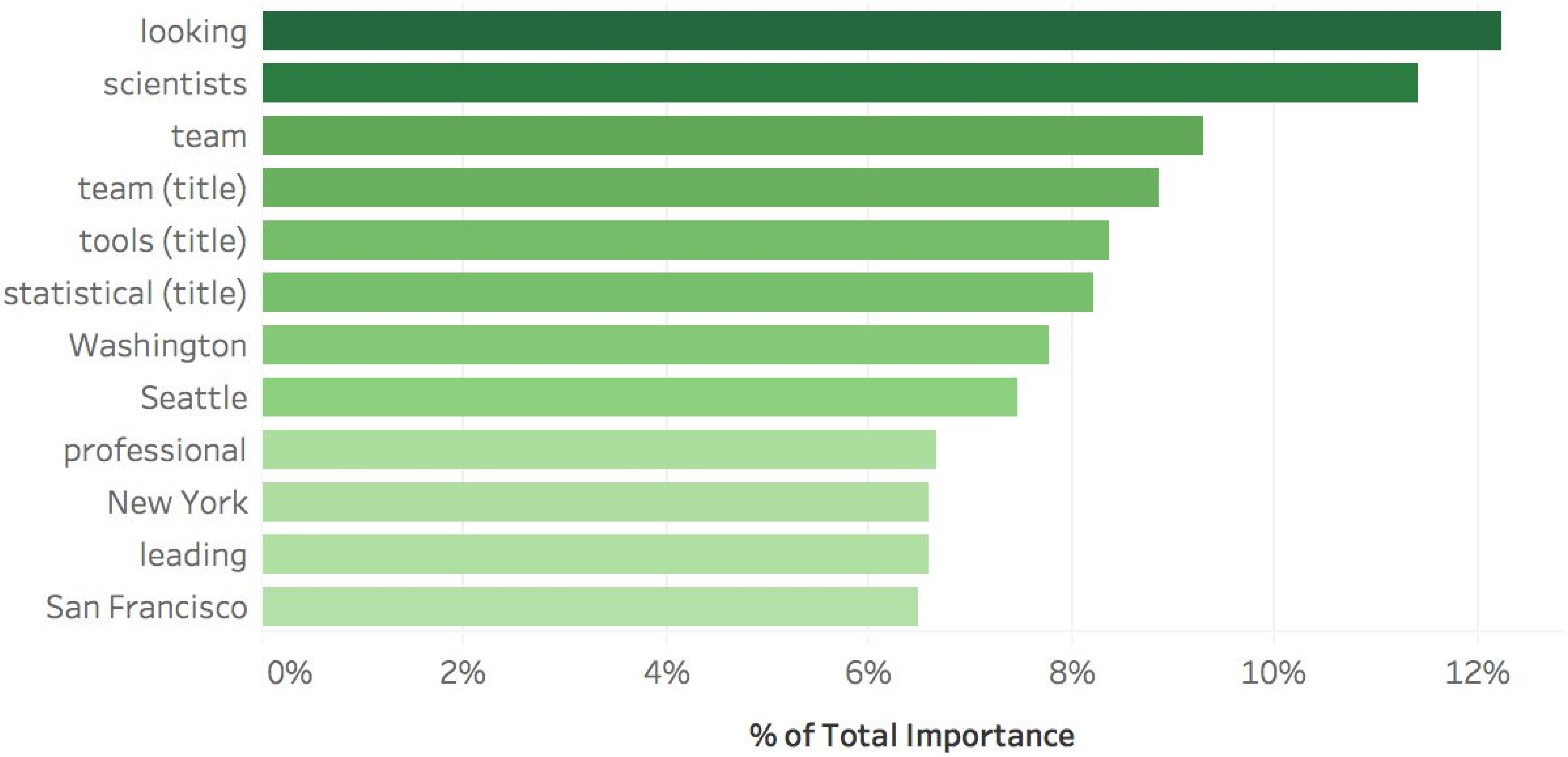


Final Models and Features

65%

Location Only

Top Final Terms



85.7%

Baseline accuracy is 50%

Text features moved accuracy from 65%

Final Terms and Location

Positive Characteristics

Looking (summary)
Scientists (summary)
Team (summary)
Team (title)
Tools (title)
Washington
Professional
New York
Leading (summary)
San Francisco

Negative Characteristics

Statistical (title)

Seattle

Other Analysis

- Classification by city
- Linear regression - predict specific salaries
- Add more classification categories
- Use strictly data science jobs in the title
- Collect qualification data

Questions