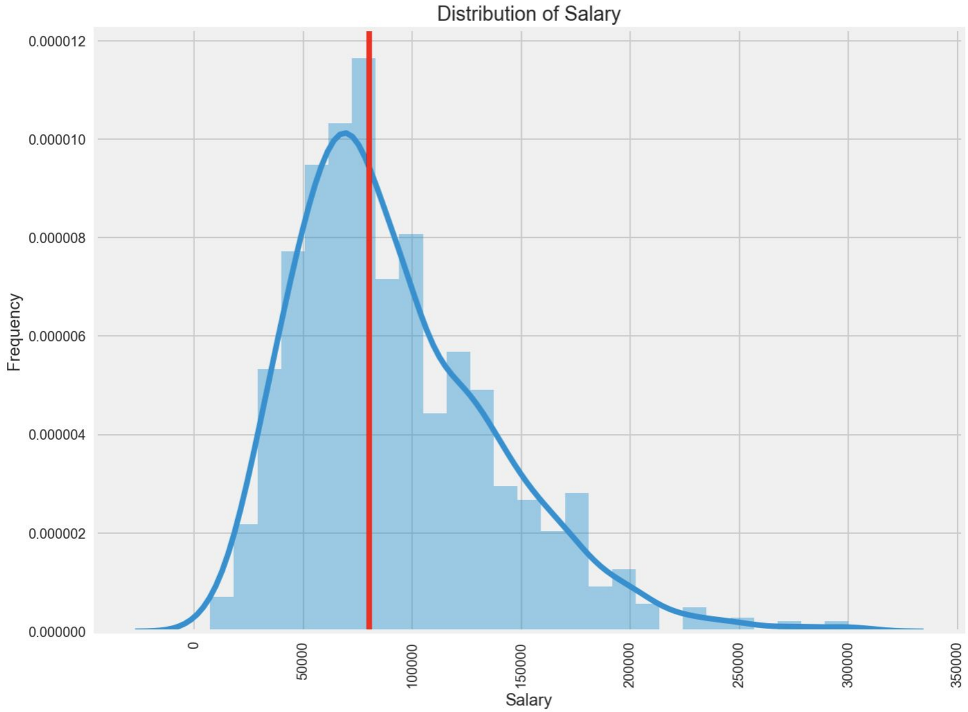
Elements of a Data Scientist's Salary

Andy Gonzalez

Building a Forest



- Data is assumed non-parametric

- Webscrape indeed.com for job title, company, salary, and location

- Random Forest model

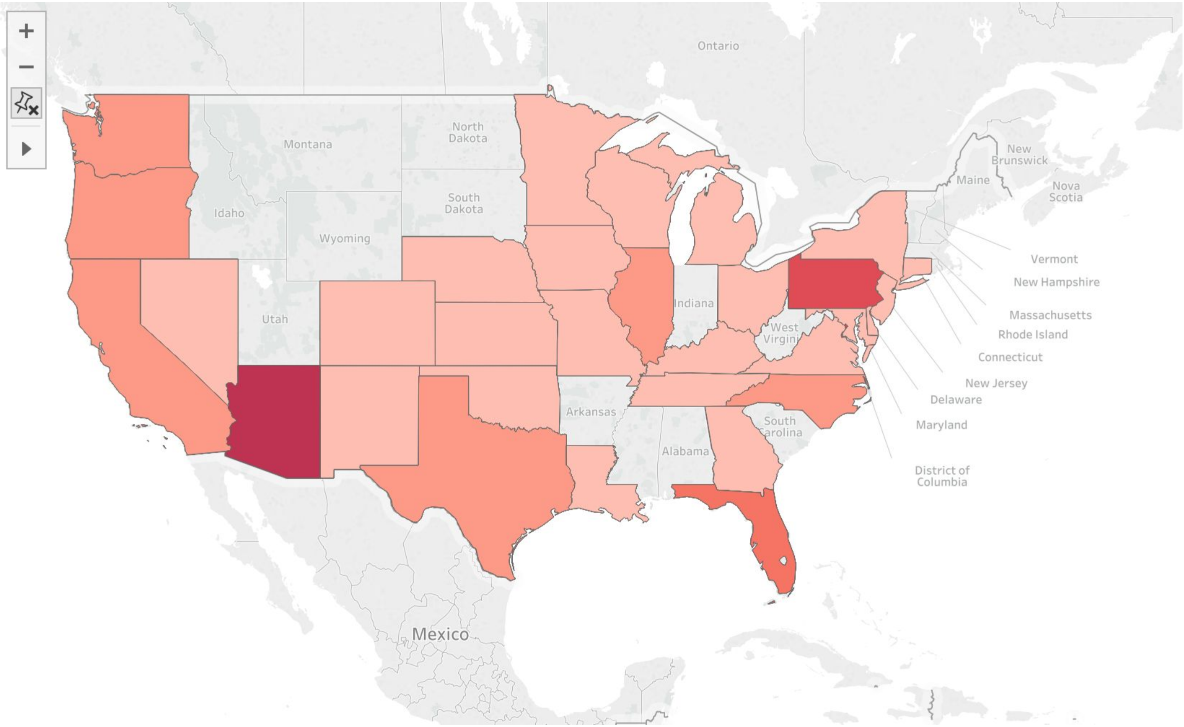
- Median as cutoff point

Importance of States as Features in Model

- Location by state

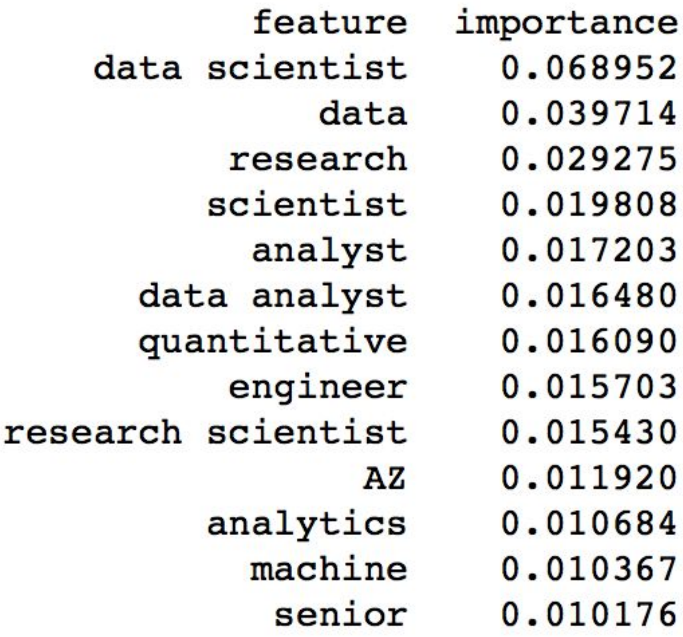
- Other features were derived from analyzing the text in the job title, assigning value by frequency(aka, Count Vectorizer, or vectorizing)

- Model tested at 89% accuracy



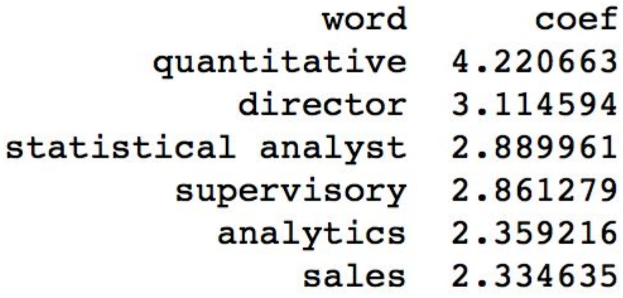
By the Numbers

Random Forest

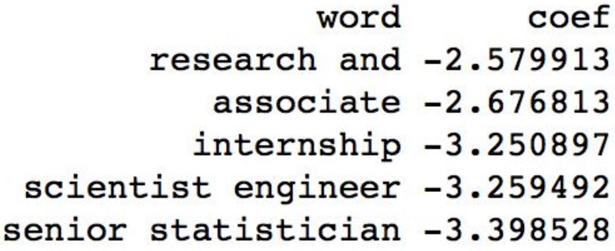


```
            feature  importance
     data scientist    0.068952
               data    0.039714
           research    0.029275
          scientist    0.019808
            analyst    0.017203
       data analyst    0.016480
       quantitative    0.016090
           engineer    0.015703
 research scientist    0.015430
                 AZ    0.011920
          analytics    0.010684
            machine    0.010367
             senior    0.010176
```

Logistic Regression
(SciKit Learn)



```
                word      coef
        quantitative  4.220663
            director  3.114594
 statistical analyst  2.889961
         supervisory  2.861279
           analytics  2.359216
               sales  2.334635
```



```
                word      coef
        research and -2.579913
           associate -2.676813
          internship -3.250897
  scientist engineer -3.259492
 senior statistician -3.398528
```

Key Takeaways

- Data derived mean salary: $92,254

- Data derived median salary: $80,395

- As far a predicting above or below median salary:
    - Important terms/characteristics: Research, analyst, quantitative, engineer, Arizona
    - Positively correlated: Quantitative, director, statistical analysis
    - Negatively correlated: Research, associate, internship