**Data Science Salary Estimator**

**Project Overview**

1. Created a tool that estimates data science salaries to help data scientists negotiate their income when they get a job.
2. Scraped over 1000 jobs descriptions from glassdoor using python and selenium
3. Engineered features from the text of each job description to quantify the value companies put on python, excel, aws and spark.
4. Optimized Linear, Lasso, and Random Forest Regression using GridsearchCV to reach the best model.
5. Built a client facing API using flask.

**Code & Resources Used**

Python Version: 3.7

Packages: pandas, numpy, sklearn, matplotlib, seaborn, flask, json, pickle

For Web Framework: pip install -r requirments.txt

Scraper GitHub: <https://github.com/arapfaik/scraping-glassdoor-selenium>

Scraper Article: <https://towardsdatascience.com/selenium-tutorial-scraping-glassdoor-com-in-10-minutes-3d0915c6d905>

**Web Scraping**

1. Job title
2. Salary Estimate
3. Job description
4. Rating
5. Company
6. Location
7. Company headquarters
8. Company size
9. Company founded date
10. Type of ownership
11. Industry
12. Sector
13. Revenue
14. Competitors

**Data Cleaning**

After scraping the data,

1. Parsed numeric data out of salary
2. Made columns for employer provided salary and hourly wages
3. Removed rows without salary
4. Parsed rating out of company text
5. Made a new column for the company state
6. Added a column for if the job was at the company’s headquarters
7. Transformed founded date into age of company
8. Made columns for if different skills were listed in the job description
9. Column for simplified job title and seniority
10. Column for description length

**EDA**



