Web Scraping Indeed Data Science Positions
Executive Summary
Aakash Tandel
July 27th 2017

Abstract: This analysis sought to understand the salary range for data scientists across the United States. Additionally, the analysis looked at which skills, titles, and locations yielded higher salaries for data scientists. The median data science salary for the US was found to be $86,711. Data was harvested from Indeed.com. The models showed that titles and descriptions containing words like "data scientist" and "machine learning" tended to have higher salaries than the median salary. Phrases like "research analyst" and "data analyst" tended to have lower salaries. The location of the job also had an effect on salary. The model showed Washington, DC tended to have higher than median salaries for data scientists.

Introduction: The analysis pulled data from Indeed.com, the job-searching site. A web scraper was built using Python. The web scraper pulled over 180,000 jobs related to data science. The web scraper searched for "data science" in the Indeed search field. The results included roles like data analyst, research analyst, statistical analyst, and data engineer along with variants of data scientist. The web scraper pulled from over 70 cities in the US. Major metropolitan areas like New York City, San Francisco, and Washington, DC gave more results than other cities. Only the jobs that contained yearly salary data were included in the model building process. Hourly jobs were dropped. The final dataset consisted of 409 jobs with complete information. Though this was less than the analysts had originally hoped for, they believed this was sufficient to build models.

Methods: The data was processed using a variety of classification models. The details of these models are not of great importance. The median salary of $86,711 was used to delineate jobs into either above the median jobs or below the median jobs. The analysts used Random Forest Models, Logistic Regression, Support Vector Machines, and XGBoost to model the data. Random Forests and XGBoost performed the best. Again, the details of these models are beyond the scope of this executive summary. The models analyzed how the title, location, and description affected which class of salary the job fell into. The models were able to correctly predict whether a job was above or below the median salary with an accuracy of 70%.

Results: The analysis found titles and descriptions containing words like "data scientist" and "machine learning" tended to have higher than median salaries. "Research analyst", "data analyst", and "statistical analyst" were titles or phrases associated with below median salaries. The location of the job had a determinative effect on salary. If a job were located in Arizona, it tended to have a below median salary. Jobs in the states of Washington, Illinois, and Massachusetts tended to have higher than median salaries. Jobs in Washington DC also tended to be higher than the median of $86,711.

Recommendations: The web scraper should be run bi-weekly for a few months in order to aggregate enough data to better model the salary data. As with many situations, more

data will only lead to more accurate models and better explanations. As a company hiring data scientists, it could be financially savvy to post job vacancies with titles such as research analyst or data analyst. These tend to be associated with lower than $86,711 salaries. If the job does not entail the use of advanced machine learning algorithms, it could be feasible for a data analyst to fill the role at a lower cost to the firm than hiring a full-fledged data scientist. Washington DC is a more expensive city in which to hire a data scientist than average so advertising a data analyst role as opposed to a data scientist role may make financial sense.