Investigation of the features of Data Scientists Jobs in the United States

Shulin Qing

October 24, 2017

Introduction

Big data are becoming ubiquitous in 21th century. Different from traditional data collected in past centuries, big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using on-hand data management tools or old-fashioned data procession applications [1]. People are trying hard to unearth patterns from big data and are enthusiastic to predict the future trend. Data scientists are the names of this group of people. They have sophisticated discipline of statistics and mathematics and rigorous training in computer science, and blend these disciplines into a harmonious and united entity. New forms of value can be extracted from a large scale of data through the insights and knowledge of data scientists. Take the recommender system as an example of the application of data science to our daily life. A great number of companies have fervidly used recommender system to promote their products according to their online customer's interests and relevant information [2]. When we browse Youtube, we may often notice that some video suggestions are showed up based on our previous watching history. The recommender system summarizes and explores information from customer's search results and uses algorithm and statistical models to predict preferences of customers. Nowadays, data scientistis jobs has been becoming one of the most popular jobs to pursue. As noted by the blogger, Christopher Watkins, from Udacity, data scientists can be seen as a hybrid of data hacker, analyst, communicator, and trusted adviser [3]. The promising perspectives attract an increasing number of people to become a member of the data scientist team. What they pay attention most are the most common skills that employers look for and the most unique skills that would make employers impressive. Moreover, understanding what types of companies employ the most data scientists could help job seekers get an overview of their future job industries and get a sense of whether the jobs would match their life goals. In this data analysis project, we performed data analysis of data scientistis jobs listed on "Glassdoor", one of most popular job search websites, and the analysis results would be used as data evidence to answer the above questions.

Method

Data Collection

We explore data science related jobs in the United States posted on glassdoor.com which is a job aggregator that updates multiple times daily. We searched for the keyword "data science" on glassdoor.com. Each page of job results have about 30 job postings and our data were collected from 30 pages. 921 jobs posted on October 8th 2017 were scraped [4]. We looped through each job listing on a page.In order to acquire the CSS for a webpage element, the Chrome SelectorGadget extension was used. By clicking on the page element that we would like the selector to match, a minimal CSS selector for the page element would be generated in a box in the bottom right of the website. We scraped job titles, company names,job locations and job descriptions on every job listing page, by specifying CSS selector ".strong", ".padRtSm", ".ib", and ".desc" respectively.The Chrome SelectorGadget extension is not effective on getting a useful CSS selector for information about industry and company size. We instead read all text lines from a job link connection, and then detected and extracted keyword related to industries and company size (See Supplemental code section 2). We scraped the information across 30 pages and collected raw data for data scientist jobs. We evaluated if each job description contains a keyword in a specified data science skill set (see supplemental

code section 2). The specified data science skills were chosen from websites including "Forbes", "Udacity" and "mastersindatascience.org" and they are Python,R,SAS,SQL,Java, Tableau, C, Perl, Excel, MATLAB,and HIVE.

The duplicated observations were removed from our raw data set. For each job posting, the following attributes were collected: company name, job title, job location specified by city and state, industry, company size and a set of data science skills as mentioned above. We identified a total number of missing cases from each variable and found that they are all less than 10%. Thus, our statistical analysis would be not likely to be biased by the missing data if they were removed. The cleaned data set for later analysis has 531 observations

For our data analysis purposes, we collapsed the categorical variables industry and job title into fewer categories respectively. The "United States Census Bureau" provides information on collapsing industry categories so 63 industry types in the data set were collapsed into 16 industry categories [5]. To identify the most unique skills that employers look for, job titles were collapsed into 6 job categories so we could find out the unique skills that employers ask for a data scientist that other similar jobs do not. For modeling purpose, we also categorized each job title based on whether it is an advanced skill level job.

Exploratory Data Analysis

Exploratory data analysis was used to summarize the most common data science skills that employers look for, and to investigate the attributes of companies that demand for data scientists. The occurences of skills and job locations were ranked and visualized by using bar charts. The industry category combined with company size for each company is visualized by using stacked bar chart, which helps us to learn the relationship between company size and industry [6].

Statistical Analysis

In order to identify the most unique skills that employers look for a data scientist, proportions of skills were compared between data scientist jobs and data analyst jobs, by using two-proportions test. The greater the difference between the proportions of a skill required by data scientist jobs and data analyst jobs is, the higher the probability that the skill is more unique for data scientists. We are also interested in investigating which factors could have significant influence on whether a job seeker would get an advanced skill level job. We used logistic regression:

$$logOdds(skill\ level) \sim skill_1 + skill_2 + ... + skill_n + company\ size + \epsilon$$

(n = 11). Then we extract and report the significant variables with p-value ≤ 0.05. The model was evaluated by cross-validation and the predicted binary response was assessed by ROC curve and the area under the curve (AUC). The analysis results would help people who look for a senior data scientist job better prepare for the job.

Results:

Skill Tag

We counted the frequency of each skill for all data scientist job postings and the top five skills that data scientist jobs look for are Python, R, SQL, Java and SAS. Among the 531 job listings, 301 (56.7%) jobs require the Python skill, 278 (52.4%) jobs require the R skill, 239 (45.1%) jobs require the SQL skill, 129 (24.3%) jobs require Java skill and 127 (23.9%) jobs require the SAS skill (Figure 1a).

The unique skills required for data scientists are Python, Java, and C (Table 1). A difference in skill proportions between data scientist jobs and data analyst jobs was tested by using 2-proportions test. Data scientists are more likely to have Python programming skill than data analysts (Data Analyst:16%,Data Scientist: 22%,p = 0.041). Data scientists are more likely to have Java programming skill than data analysts (Data Analyst:3.2%,Data Scientist: 9.3%,p = 0.004) and C language is more uniquely reqiured by data scientists than data analysts (Data Analyst:2.7%,Data Scientist: 6.5%,p = 0.032).

Table 1: **Skill Differences between Data Scientists and Data Analysts.** Python, Java, and C language are more uniquely reqiured by data scientists than data analysts

	Data Analyst	Data Scientist	P Value	95% Conf Int
python	0.160	0.220	0.041	(-1,-0.007)
R	0.165	0.205	0.125	(-1,0.013)
SAS	0.160	0.083	0.999	(-1,0.127)
SQL	0.229	0.138	0.999	(-1,0.148)
JAVA	0.032	0.093	0.004	(-1,-0.031)
Tableau	0.059	0.045	0.720	(-1,0.047)
C	0.027	0.065	0.032	(-1,-0.011)
Perl	0.016	0.014	0.500	(-1,0.021)
Excel	0.080	0.032	0.998	(-1,0.085)
MATLAB	0.027	0.048	0.140	(-1,0.005)
HIVE	0.048	0.059	0.338	(-1,0.021)

Location

Figure 1b suggests that the top four cities that have the most number of data science jobs are New York (NY), San Francisco (CA), Chicago (IL) and Cambridge (MA). These cities are nationwide job centers which emcompass lots of tech companies.

Industry and Company Size

The stacked bar chart of industry job categories is shown in Figure 1c. The bars are stacked to different colors which represents different company sizes. Each colored rectangle represents a combination of industry and company size. As shown in the bar chart, the top four industries to find a job in data science are "Professional, Scientific and Technical Services", "Information", "Manufacturing" and "Finance and Insurance". As defined by the United States Census Burea, the industry "Professional, Scientific and Technical Services" includes industry subcategories: legal services, accounting, engineering related services, computer systems design, and scientific and technical consulting services, etc. This industry provides 133 (25.0%) data science related jobs. Information industry includes publishing, motion picture, radio and television broadcasting, and telecommunication services, etc. It provides 131 (24.7%) data science related jobs. Manufacturing industry provides 68 (12.8%) data science related jobs and finance and insurance industry provides 58 (10.9%) data science related jobs. In professional,scientific and technical services industry, most companies that demand for data scientists are large companies that have more than 250 empoyees. Same patterns can also be seen in other industries that have observations more than 10, except administrative and support industry. According to international standards, greater than or equal to 250 employees is taken as an indicator that it is a large business and less than 50 employees is taken as an indicator that it is a small business [7].In administrative and support industry, more than a half of companies that are in need of data scientists are small companies as shown by the red color. Too few observations (less than 10) in the industries may bias our analysis so we exclude them from our discussion.

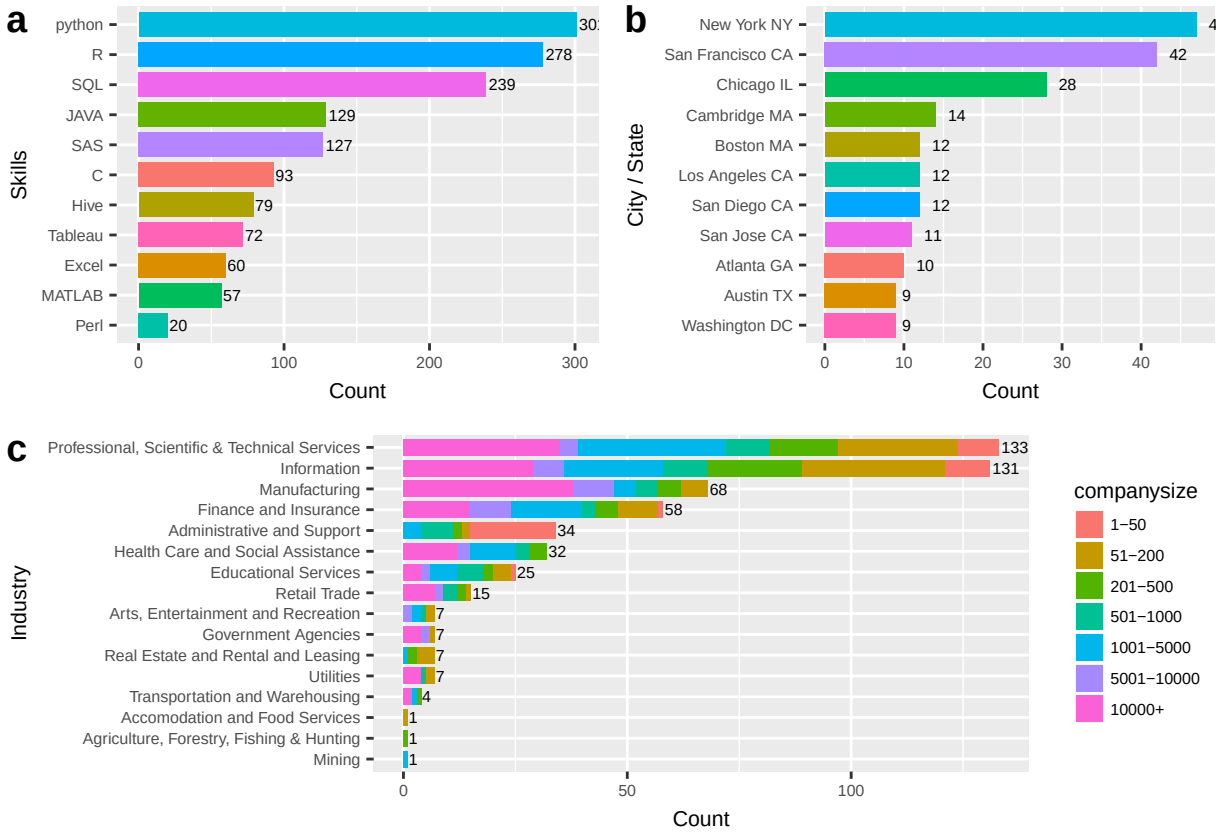

Figure 1: **Frequency Tables of Job Posting Features.** (a) A bar chart of data science skills. The top five skills that data scientist jobs look for are Python, R, SQL, Java and SAS. (b) A bar chart of job locations. The top four cities that have the most number of data science jobs are New York, San Francisco, Chicago and Cambridge. (c) A stacked bar chart of industry job categories stacked with different colors representing different company sizes. The top four industries to find a job in data science are Professional, Scientific and Technical Services, Information, Manufacturing and Finance and Insurance. In most industries except Administrative and Support industry, a very high percentage of companies that demand for data scientists are large companies that have more than 250 empoyees.

Factors Impacting Job Skill Level

We observed Python,SQL and Java have statistically significant ($p < 0.05$) association with a job skill level and R programming has moderately statistically significant ($p = 0.08$) association with a job skill level. There is no significant association between a job skill level and a company size even though we might think that a larger company with more sections may provide more advanced level data science jobs. The odds of getting an advanced level data science job for people who are proficient in R programming is 1.94 times the odds of getting an advanced level data science job for people who do not have R programming skill with 95% CI (0.917,4.10). The odds for people who have Python skill getting an advanced level data science job is 0.41 times the odds for people who do not know Python with 95% CI (0.20,0.86). The odds for people who have SQL skill getting an advanced level data science job is 1.90 times the odds for people who do not know SQL with 95% CI (1.01,3.57), and the odds for people who have Java skill getting an advanced level data science job is 0.35 times the odds for people who do not know Java with 95% CI (0.13,0.96). Thus, we could expect that people with R programming and SQL skill are more likely to get an advanced level jobs such as directors and managers. We performed cross-validation and used ROC curve and AUC to examine the model fitting (See Supplemental Figure 1). The AUC is 0.51, which means our model does not have much prediction power.

Discussion:

Our analysis provides some insights to job seekers who are interested in data science jobs. To become qualified candidates for data science jobs, people should learn Python, R programming, SQL, JAVA and SAS since they are the top five skills that data science jobs look for as suggested by our exploratory analysis. Not only Python and Java are the most common skills a data scientist should be equipped with, they are also the most unique skills employers ask for data scientists. C programming is also more uniquely required for data scientist jobs compared to other similar jobs. Job seekers who aim at data science jobs want to know the characteristics of companies that hire data scientists. A large number of companies that employ data scientists are located in the west coast and northeast region such as San Francisco and New York, which are the technology hub of the country. The top 10 cities that hire most data scientists are all very large cities so it is more likely to get a data scientists jobs in a metropolitan city which provides more opportunities. It is also more likely to find a data science job in “Professional, Scientific and Technical Services”, “Information”, “Manufacturing” and “Finance and Insurance” industries. Most of them are high-tech industries and are more involved in big data. Generally, in most industries, large companies that have more than 250 empoyees are more likely to hire data scientists. This may due to the fact that large scale companies produce a high amount of data every day and require a great number of data scientists to deal with the data and get insights.

Our data were scraped from glassdoor.com and thus the results may not apply to all data science job postings on the internet. The frequencies of skills and features of companies that recruit data scientists may be biased because of the single data source. The glassdoor.com charges companies for posting job information on its website so we only collected data from companies that would like to pay for the job posting and advertising. These companies on glassdoor.com thus do not represent all companies in the United States. In order to mitigate the bias, we would further explore data science jobs in multiple websites such as linkedin, indeed and Monster.

References

1. Press, Gill "12 Big Data Definitions: What's Yours?", Forbes Page. URL: https://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/#199010c513ae. Accessed 10/10/2017.
2. Analytics Vidhya "13 Amazing Applications / Uses of Data Science Today" Page. URL: https://www.analyticsvidhya.com/blog/2015/09/applications-data-science/. Accessed 10/10/2017.
3. Watkins, Christopher "Hottest Jobs in 2016 #3 Data Scientist", Udacity Page. URL: https://blog.udacity.com/2016/01/hottest-jobs-in-2016-3-data-scientist.html. Accessed 10/10/2017.
4. Hadley Wickham (2016). rvest: Easily Harvest (Scrape) Web Pages. R package version 0.3.2. https://CRAN.R-project.org/package=rvest
5. United States Census Bureau, "Annual Capital Expenditures Survey (ACES)" Page. URL: https://www.census.gov/programs-surveys/aces/information/iccl.html. Accessed 10/10/2017.
6. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.
7. Support Site, "What is the definition of Small, Medium and Large business within the Observatory?" Page. URL: http://support.spikescavell.com/faqs/what-is-the-definition-of-small-medium-and-large-business-within-the-observatory/
8. Yihui Xie (2017). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.16.
9. Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963
10. Hadley Wickham, Romain Francois, Lionel Henry and Kirill Müller (2017). dplyr: A Grammar of Data Manipulation. R package version 0.7.4. https://CRAN.R-project.org/package=dplyr
11. Hadley Wickham (2017). tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions. R package version 0.6.1. https://CRAN.R-project.org/package=tidyr
12. Yihui Xie (2017). printr: Automatically Print R Objects to Appropriate Formats According to the 'knitr' Output Format. R package version 0.1. https://CRAN.R-project.org/package=printr
13. Kun Ren and Kenton Russell (2016). formattable: Create 'Formattable' Data Structures. R package version 0.2.0.1. https://CRAN.R-project.org/package=formattable
14. Sing T, Sander O, Beerenwinkel N and Lengauer T (2005). "ROCR: visualizing classifier performance in R." *Bioinformatics*, *21*(20), pp. 7881. <URL: http://rocr.bioinf.mpi-sb.mpg.de>.
15. Alboukadel Kassambara (2017). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.1.5. https://CRAN.R-project.org/package=ggpubr
16. Hadley Wickham (2016). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.1.0. https://CRAN.R-project.org/package=stringr