Analysis of American Data Scientist Jobs: Skills, Salary and Satisfaction

Bingkai Wang

October 7, 2017

1 Introduction

2 Data

Job data was scraped from 'Glassdoor'(www.glassdoor.com) with title "data scientist" using `rvest` package in `r`. Glassdoor is a job search engine and outperforms others in the sense that it contains reviews of employees to the companies and positions. From this website, 990 job posted in September 30th 2017 were scraped. (Glassdoor API is not accessible for students and only 1000 jobs can be viewed by normal users.)

For each job posting, the following features were collected: company name, rating, salary, location (city, state), company size, industry and full text of job description. Among these variables, rating is the overall rating of all ratings of all time. Salary is preprocessed to be the mean of the salary range provided in the website. 722 out of 990 job postings contain no missing value. For explortary analysis, missing values are ommited. For inferencial analysis, for simplicity, only complete cases are used.

3 Methods

3.1 Identification of 'skills' from job descriptions

Our first goal is to identify phrases that represent 'skills'. The traditional way for achieving this is to tokenize the text (partition it into a vector of words), count the frequency of each word and remove the meaningless 'stop words', such as 'the', 'we' and 'of'. A set of common stop words can be find *here*. However, this method fails to work in this case for two reasons. First, the common set of stop words does not contain enough words for identifying skills in job description. For example, words like 'area', 'opportunity', 'career' have much larger frequency than skill words like 'python', 'statistics', but cannot be filtered out by stop words. Furthermore, such type of words is so many to filter manually. Second, skills are often appeared as phrases rather than single words. When performing tokenization, 'machine learning', 'data mining' and etc will be seperated into 'machine', 'learning', 'data' and 'mining', which no longer have explicit meaning as skills.

To overcome these two difficulties, we provided a method a feasible way to extract 'skills' from job description. Our method is based on the following observation: in job descriptions,

skills are usually listed, starting with ":" and seperated by ",", "and" or "or". Hence, if we partition the text by these symbols, skill phrases will be indentified as a whole and appear in high frequency since other noise words will remain in long and unique segments. For example, in the following sentence of some job description,

> "BACKGROUND/EXPERIENCE: Demonstrates proficiency in most areas of mathematical analysis methods, machine learning, statistical analysis, and predictive modeling and in-depth specialization in some areas."

we will identfiy 'machine learning' and 'statistical analysis'. As we collect more job descriptions, such skill phrases will appear many times and be distinguishable. Finally we use these skills as key words to search in each job description and get the ultimate count of skill words. The final step is necessary because not all skills are listed in each job description.

As a result, we identify skill phrases with these steps:

- **Step 1** tokenize the text of job descrption by seperator ":", ",", ".", "and" and "or";
- **Step 2** extract the tokens with 3 or fewer words;
- **Step 3** count the frequency of each token and remove those appearing 50 times or less (58 phrases remained);
- **Step 4** filter out those irrelavent words manually, such as "religion" and "gender" (32 phrases remained);
- **Step 5** Use the remained phrases as keyword, search them in every job description and produce an indicator for each keyword in each job description.

If we assume that each skill appears in each job posting independently with probability $p \geq 0.07$, then by binomial distirbution, the skill will remain in the final list with probability at least 0.99.

3.2 Linear models for salary and employees' satisfaction

Results

3.1 Explortory analysis

3.2 Factors impacting salary and satisfaction

Discussion