Analysis of US Data Science Job Market In 2018*

with Text Mining and Classification using Spark

Xudong Peng†
Department of Statistics and Actuarial Science
University of Waterloo
Waterloo ON Canada
x46peng@uwaterloo.ca

Daoyi Chen
Department of Statistics and Actuarial Science
University of Waterloo
Waterloo ON Canada
d89chen@uwaterloo.ca

ABSTRACT

As prospective data science graduates, one of our most questions would be what kind of data science-related jobs we can do after graduation and what qualifications we are expected to have. As CS631 project of type 4, we performed data analysis on the data science job dataset crawled from Indeed.com and implemented text mining and classification techniques under spark environment.

This project consists of solving two main tasks. The first one is to figure out how data science job distribute in the US and the second one is to classify "Data Engineer", "Data Scientist " and "Data Analyst" based on job description (predictors) and job type (label) via different machine learning algorithms, and how you could prepare (degree, major, programming language) for getting a data science job in the US. Specifically, the sections of the current report are as follows: dataset description, distribution of data science jobs, text mining of key qualifications, classification on three types of data jobs, conclusion and future work. The tools and concepts used in this project include pyspark, spark dataframe and spark rdd, and we implemented text mining techniques such as term frequency, keyword extraction, and pair searching.

KEYWORDS

Text Mining, Word Count, Pair Search, Pyspark, Spark Dataframe, Spark RDD, Data Visualization, Machine Learning, Logistic Regression, Random Forest, Naive Bayes, Text Classification.

1. Dataset Description

The dataset is originally from Indeed website on August 2018 [1] and is obtained by Silvia Lu [2] through web scraping. This dataset has a total of 6953 job postings and each posting including job position company, number of reviews, description, and location. Specifically, description contains over 1.7 million text-based information detailing company introduction, job duties, job qualification and application procedures. The data was initially cleaned and imputed to ensure there are no missing values. The most time-consuming part of this project is to preprocess data, which includes removing unnecessary characters and converting into proper formats for text mining.

2. Text Mining of Job Distribution

This section includes parts of basic term frequency, data visualization of cities/states with most jobs, companies with most jobs to give readers a general idea on how US data science job market looks like in 2018.

2.1. Term Frequency

By simply counting the words of the whole dataset after tokenization and stop words removal, we have top words as follow:

```
[('data', 43713),
 ('experience', 32769),
 ('work', 19560),
 ('team', 16916),
 ('skills', 13150),
 ('development', 13018),
 ('business', 12239),
 ('learning', 10577),
 ('science', 9704),
 ('analysis', 9241),…]
```

It looks like simply word counting does not work well and we did not get much accurate information, but we still see some important keywords like "data", "analysis", "science" and "analytics". We would continue digging into these data science jobs based on degree, major, top wanted qualifications, and top wanted programming language in later part of the report. First, we targeted our keywords on the investigation of job location(city/state).

2.2. Cities and States with Most Jobs

Where is the data science hub in the United States? We want to find the cities that have the most data scientist jobs. Our

* Article Title Analysis of US Data Science Job Market In 2018
† Author Footnote Xudong Peng, Daoyi Chen

DATA SCIENCE'19, April, 2019, Waterloo, ON Canada

assumption is that west coast cities such as San Francisco or Seattle would be the top one city with the most data science jobs since there are a lot of high-tech companies.

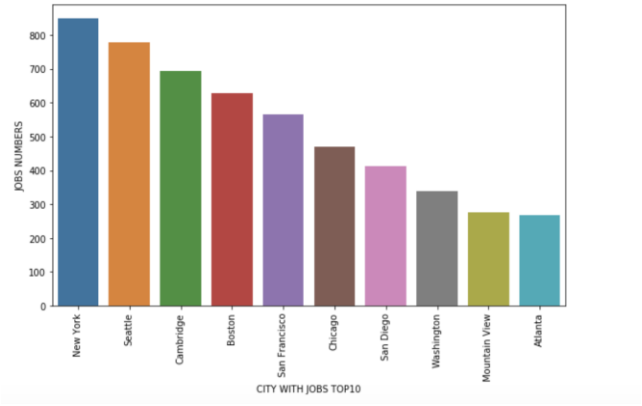

Figure 1: Cities with Most Data Science Jobs Top 10

Surprisingly in Figure 1, New York has the most data science jobs, and Seattle is following as the second place. It is also noticeable that Atlanta, as a rising star, takes the 10th place. One interesting thing is that Mountain view here is at the ninth place among these big cities. Mountain View is home to many high technology companies and is one of the major cities that make up Silicon Valley. It is reasonable to imagine that if counting jobs based on areas, Silicon Valley should be the very top area that has most data. Also, we can see that Cambridge is the bronze medallist among these big names cities. The reason could be that Cambridge is home to top post-secondary institutions such as Harvard University and the Massachusetts Institute of Technology (MIT) and, therefore, Cambridge has attracted a lot of companies that want to hire top talents directly.

Figure 2 shows data science job numbers based on states. Not surprisingly, San Francisco, Mountain View, and San Diego, make California as the state which has the most data science jobs.

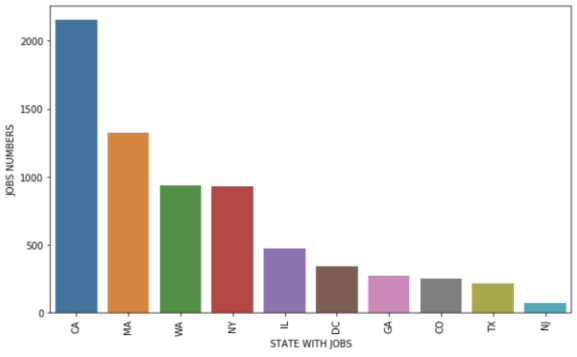

Figure 2: States with Most Data Science Jobs Top 10

We can see New York state ranks as the fourth place with more than 900 job opportunities, which has a similar size as the number of data science jobs in New York City. Also, for the state of Illinois, it also has the same scale of data science jobs as in Chicago. Obviously, these states only have one major data hub, compared to those states with more potential choices of cities, such as the state of California and the state of Washington. For people who are looking for data science jobs and moving to one of those data hubs, the state of California and the state of Washington would offer more opportunities and flexibility.

Moreover, New Jersey state takes 10th place without any cities in top places, actually, one of the main cities in New Jersey state, the Jersey City is the 24th place as shown in Figure 3. We can see that data science jobs do distribute extremely unevenly across cities and states in the US.

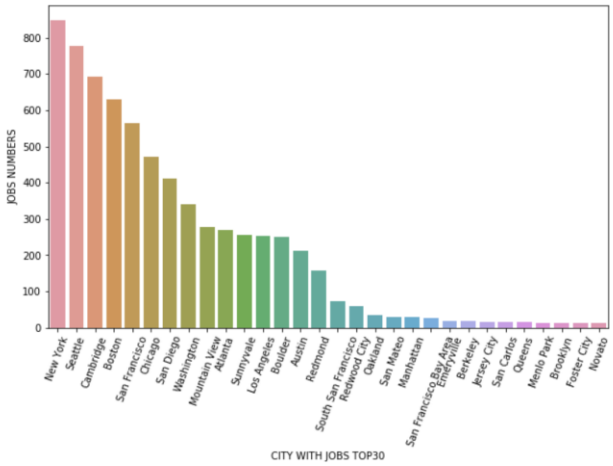

Figure 3: Cities with Most Data Science Jobs Top 30

2.3 Companies with Most Jobs

Then we also target companies that have most data science opportunities. By looking into the top ten companies that hire most data science people, big high-tech names like Amazon, Google, Microsoft, Facebook are in the top-ten list in anticipation, with the exception that Facebook is in the last place of top ten list.

However, Ball Aerospace takes second place for having most data science jobs. It is an American manufacturer of spacecraft, components, and instruments for national defense, civil space, and commercial space applications. Also, the 4th and 5th places are two health research institutions NYU Langone Health and Fred Hutchinson Cancer Research Center. Furthermore, the company Lab126, ranking at the 8th place, is an American research and development and computer hardware company owned by Amazon.com. For people who are looking for data science opportunities, they can not only keep eyes on the high-tech industry but also more in medical health, research, and financial service (like KPMG) industries.

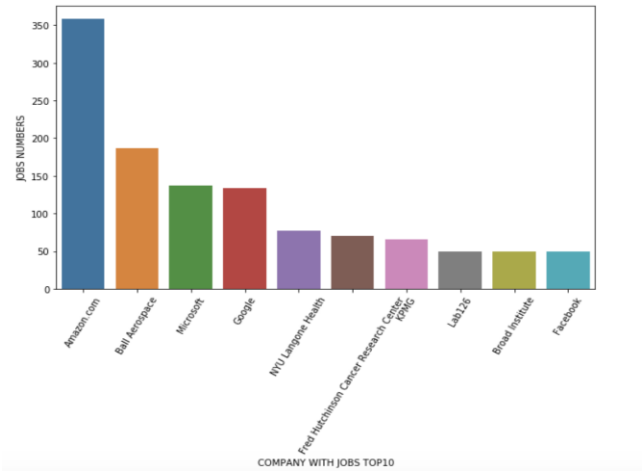

Figure 4: Companies with Most Opportunities Top 10

3. Classification

3.1. Data Exploration

In this section, we added a number of new features based on position type and company size and compared their differences in terms of skill, major, degree and tool.

First, the positions were grouped into three categories (data scientist, data analyst, and engineer) based on their keywords in job titles, while the remaining jobs which are ambiguous were classified as others. In the meantime, we also created a new column called company size which is based on the number of reviews a company received. Figure 5 describes the position distribution in different sizes of companies. We see that data scientists occupy the most part of positions in small companies, and all companies have a similar number of engineers and data analysts.

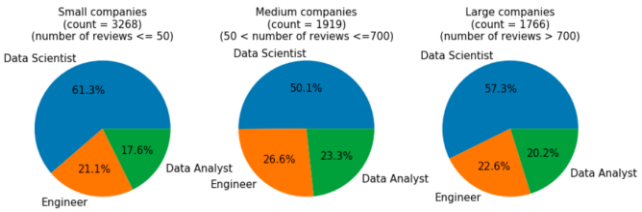

Figure 5: Position Distribution in Different Size of Companies

Next, we studied their distinct job characteristics by comparing the job descriptions of the three position types. Figure 6 shows the degree requirements for the three types of positions. It can be seen that data scientists require a higher degree on average. People with Bachelor's and Master's degrees usually become Data Analyst and Engineer. For major preference shown in Figure 7, engineer positions prefer hiring candidates with a major in Computer Science. However, data scientists and analysts do not have a strict requirement for majors. Quantitative majors such as statistics and mathematics are generally acceptable to employers. It is recommended that candidates with other degrees such as chemistry and biology should aim for a higher degree if they want to work in data science field.

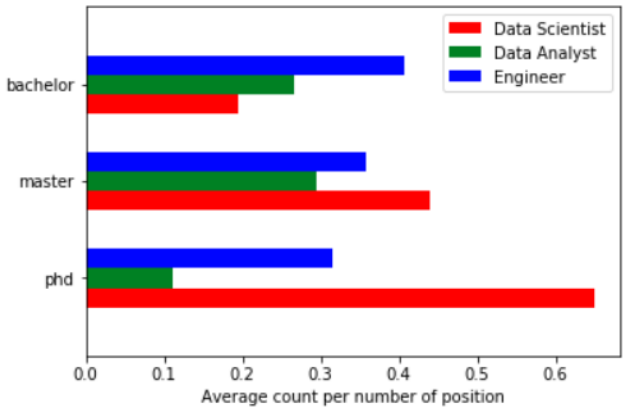

Figure 6: Degree Requirements for the Three Job Types

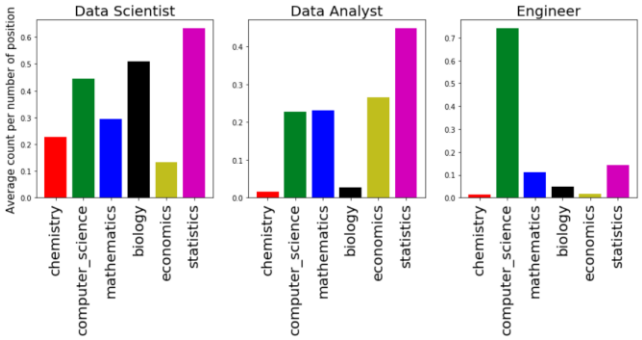

Figure 7: Major Difference in the Three Job Types

Subsequently, we investigated the effect of skill sets on the hiring process of the different job types. We can see that three distinct skills are needed by different positions in Figure 8. For example, artificial intelligence is most required in order to become an engineer, while data analyst needs strong skills in data analysis and visualization. Furthermore, companies prefer data scientists with solid machine learning, modeling, and deep learning skills. Figure 9 shows the average occurrence of data science tools in job descriptions categorized by the three job types. We can see that the most desirable tools for a data scientist are Python, R, SQL. In contrast, some programming languages related to distributed systems such as Spark, Java, Hadoop are more important for engineers. Data analysts generally prefer to use Excel the most.

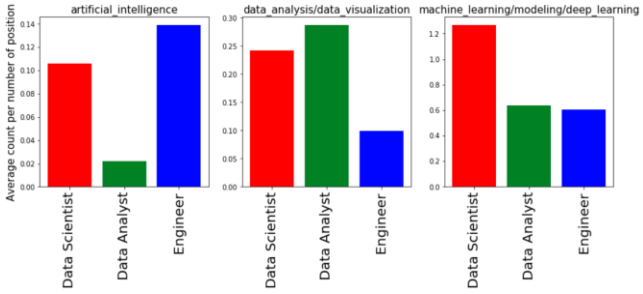

Figure 8: Skill Requirements for the Three Job Types

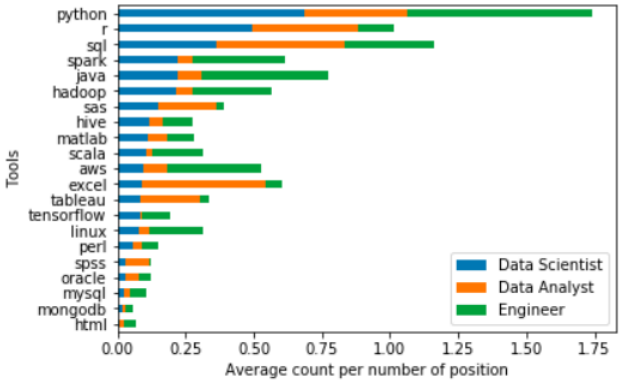

Figure 9: Tool requirements for the Three Job Types

3.2. Machine Learning Classification

Finally, we investigated the possibility of using job description to classify position type. Before model training, all information directly related to position type and common stop words were removed from the job description, which is followed by word tokenization. Both HashingTF and CountVectorizer were used to generate the term frequency vectors. Different machine learning algorithms were implemented to tackle the current multiclass text classification problem. Furthermore, the dataset was split into 70% for training and 30% for testing.

The prediction accuracy for different models is listed in Table 1. It can be seen that the random forest algorithm performs poorly as it is not suitable for the current high-dimensional sparse data. After some parameter tuning, the best prediction accuracy is 0.8230 achieved by using Logistic Regression with Count Vector features.

Table 1: Prediction Accuracy for Different Models

Models	**Prediction Accuracy**
Logistic Regression with Count Vector	0.8230
Logistic Regression with TF-IDF	0.8004
Naive Bayes	0.7811
Random Forest	0.4914

It is estimated that the prediction accuracy is limited by the fact that employers do not have a clear idea of what type of position they need. Businesses need some time to gain experience in this fast-growing industry and adapt to these new job types. Therefore, job hunters are recommended to read job descriptions carefully and not to be deceived by job titles.

4. Conclusion

From our analysis, we know the state of California and Washington have the most data science job opportunities, however, New York City is the city with most data science jobs. For current job hunters, holding a higher degree as well as proficiency in Python, R and SQL programming would improve chances getting a data science job in the US. Also, job hunters should need to pay attention to opportunities in various industries instead of the only high-tech industry. In the future, it is possible that more and more industries would need data science talents. Yet, even for data science jobs, they have different requirements under different circumstances. Since employers do not have clear ideas about what kinds of talents they want, job hunters should read carefully about job descriptions to find a good fit.

5. Future Work

Data science is an emerging industry and the threshold requirements of this industry are constantly changing with time, if possible, we can streamline web crawling data from Indeed website and make a real-time analysis of the updated data science jobs dataset, which can guide people who are looking for next step in the data science industry. Specifically, we can also develop and optimize our current classification algorithm to automate and accurately classify the wanted data job type for job hunters, as well as desired cities and states. Hopefully, this analytical project can shed some light on people who want to move the next career stage in the data science industry.

REFERENCES

[1] Data Scientist Job Market in the U.S. Available at: https://www.kaggle.com/sl6149/data-scientist-job-market-in-the-us/

[2] Who gets hired? An outlook of the U.S. Data Scientist Job Market in 2018. Available at: https://nycdatascience.com/blog/student-works/who-gets-hired-an-outlook-of-the-u-s-data-scientist-job-market-in-2018/