Proposal: How to Get a Job in Data Science Using Data Science

Melanie Desroches
Department of Statistics
University of Connecticut

October 27, 2024

Introduction As a senior in college, it is only natural that I have been considering what my career will look like post-graduation. As a result, I have been looking at job applications to see what employers are looking for in candidates and if I qualify. It is very easy to get overwhelmed by all the skills and qualifications that you need to have in order to get the job that you want. In this project, I'll explore the different types of data science roles by clustering job postings based on skill requirements and job types. Using a dataset that consists of data science related job postings from Glassdoor, the goal is to identify natural groupings, such as data engineering, data analysis, and machine learning roles, based on the skills each type of job requires. Hopefully, this project can give insight into what skills job-searchers in the realm of data science should have based on the type of job that they want.

Specific Aims The goal of this project is to identify what factors employers are looking for in a candidate for a data science job. There are different types of jobs in the realm of data science, such as data engineering, data analysis, machine learning engineering, etc. Based on the different types of roles, what skills are the most prominent? For example, do data engineering roles tend to require AWS knowledge? Python? Tableau? Based on the clusters, the aim is to uncover each "profile" and discuss which skill combinations are common. These profiles will illustrate the typical skill sets needed across different data science job types and levels.

Data The data set I will be using is based on job postings on Glassdoor. The data can be found on Kaggle. The dataset already has some initial cleaning done. The columns that will be of the most interest to me are Job Description (which is the full description on the job posting), python, excel, hadoop, spark, aws, tableau, big data (which are all boolean variables used to determine if it is a relevant skill to the job), and information about the company that is hiring (name, industry, etc). There are 660 postings prior to cleaning. It is possible that there may be duplicates so this may not be entirely accurate. There are 27 columns prior to cleaning and feature engineering. Columns may be removed or generated based on the project needs.

Research Design and Methods To perform my analysis, I will be apply a clustering algorithm such as K-means or Hierarchical Clustering. The first step will be processing and cleaning the dataset. This will involve looking for any missing values, removing any jobs that seem irrelevant (meaning they do not involve data science), and dropping unnecessary columns. This step will also involve some feature engineering. Using the job descriptions provided, skills can be pulled and turned into new columns. Once the data is prepared, the appropriate clustering algorithm can be applied. Currenlty, I am considering using either K-Means Clustering, DBSCAN, or Hierarchical clustering. Which clustering algorithm I use will depend on the shape of the data. Based on the results, data visualizations can then be generated in order to help understand the clusters better. Finally, the results can be interpeted.

Discussion The most challengings parts of the project will be selecting the best clustering technique. When picking the algorithm, I need to consider the size and shape of the data and clusters. I also need to be mindful of how the different clustering techniques may influence how the results can be interpeted. There are some limitations with respect to the data set. Since the data is being sourced from Kaggle, instead of collecting it myself, I am trusting that the owners of the dataset have collected accurate information. I am not sure when the data was collected. If it was from a while ago, it is possible that the job market has since changed and the data is no longer relevant. Also, I am operating under the assumption that all of the job postings are legitamate and are not scam posts. If I am truely unsure about the quality of the data set, I could always try to scrape the information myself.

References